7-2021

Knowledge Discovery from Complex Event Time Data with Covariates

Samira Karimi
University of Arkansas, Fayetteville

Knowledge Discovery from Complex Event Time Data with Covariates

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Engineering with a concentration in Industrial Engineering

by

Samira Karimi
Sharif University of Technology
Bachelor of Science in Industrial Engineering, 2014
Sharif University of Technology
Master of Science in Industrial Engineering, 2016

July 2021
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Haitao Liao, Ph.D.
Dissertation Director

Edward A. Pohl, Ph.D.
Committee Member

Xiao Liu, Ph.D.
Committee Member

Avishek Chakraborty, Ph.D.
Committee Member

Abstract

In particular engineering applications, such as reliability engineering, complex types of data are encountered which require novel methods of statistical analysis. Handling covariates properly while managing the missing values is a challenging task. These type of issues happen frequently in reliability data analysis. Specifically, accelerated life testing (ALT) data are usually conducted by exposing test units of a product to severer-than-normal conditions to expedite the failure process. The resulting lifetime and/or censoring data are often modeled by a probability distribution along with a life-stress relationship. However, if the probability distribution and life-stress relationship selected cannot adequately describe the underlying failure process, the resulting reliability prediction will be misleading. To seek new mathematical and statistical tools to facilitate the modeling of such data, a critical question to be asked is: Can we find a family of versatile probability distributions along with a general life-stress relationship to model complex lifetime data with covariates? In this dissertation, a more general method is proposed for modeling lifetime data with covariates. Reliability estimation based on complete failure-time data or failure-time data with certain types of censoring has been extensively studied in statistics and engineering. However, the actual failure times of individual components are usually unavailable in many applications. Instead, only aggregate failure-time data are collected by actual users due to technical and/or economic reasons. When dealing with such data for reliability estimation, practitioners often face challenges of selecting the underlying failure-time distributions and the corresponding statistical inference methods.

So far, only the Exponential, Normal, Gamma and Inverse Gaussian (IG) distributions have been used in analyzing aggregate failure-time data because these distributions have closed-form expressions for such data. However, the limited choices of probability distributions cannot satisfy extensive needs in a variety of engineering applications. Phase-type (PH) distributions are robust and flexible in modeling failure-time data as they can mimic a large collection of probability distributions of nonnegative random variables arbitrarily closely by adjusting the model structures. In this paper, PH distributions are utilized, for the first time, in reliability estimation based on

aggregate failure-time data. To this end, a maximum likelihood estimation (MLE) method and a Bayesian alternative are developed. For the MLE method, an expectation-maximization (EM) algorithm is developed to estimate the model parameters, and the corresponding Fisher information is used to construct the confidence intervals for the quantities of interest. For the Bayesian method, a procedure for performing point and interval estimation is also introduced. Several numerical examples show that the proposed PH-based reliability estimation methods are quite flexible and alleviate the burden of selecting a probability distribution when the underlying failure-time distribution is general or even unknown.

Keywords: Aggregate failure-time data; Phase-type distributions; Maximum likelihood estimation; Bayesian method.

Contents

List of Figures

List of Tables

1 Introduction

In reliability data analysis, depending on the type of product and the data collection method, various types of data may be recorded. The most straight-forward data type is failure time data from individual components in normal and equal conditions. However, this ideal type of data is rarely accessible. Important other cases of available reliability lifetime data include censored data, accelerated life testing (ALT) data, failure-censored aggregate data and time-censored aggregate data. These types of data or their combinations, while abundantly available, may raise challenges during the statistical inference and data analysis procedure. A remarkable challenge in this regard, is that not all distributions have the mathematical potential to model these types of data. To solve this issue, robust methods are provided in this work to eliminate the need for model selection and compensate for the other distributions that are not capable of modeling these types of data. For this purpose, continuous Phase-type distribution is utilized as a flexible distribution that can mimic any nonnegative distributions.

A continuous Phase-type (PH) distribution describes the time to absorption of a continuous-time Markov chain (CTMC) defined on a finite-state space. As the set of PH distributions is dense in the set of all nonnegative distribution, almost any nonnegative distribution can be well represented by a PH distribution in the sense that their first three moments agree. PH distributions have been vastly applied to many statistical analysis fields including queuing theory, healthcare problems and risk analysis. However, the amount of PH distribution usage in degradation and survival reliability analysis has been relatively small, leaving a gap in the literature.

Substantial reliability life data types include ALT data and field data. ALT data which is collected in laboratory, is a significantly practical, although expensive, source of reliability data for the companies. In this type of data collection, the products are exposed to different levels of environmental stresses, such as temperature, humidity, radiation and voltage, to expedite the failure process. In the literature, models have been created to analyze the data from ALT, using exponential, Weibull and lognormal distributions. Nonetheless, it can be priceless to create a robust

model that can eliminate the model selection process. Another significant data source is field data. The advantage of field data is its authenticity, as it is collected in real usage conditions, and being economical, as no extra resource or product is required. An important source of reliability data belongs to Reliability Information Analysis Center (RIAC) which is designed to collect reliability data from U.S. DoD organizations. According to Coit and Jin (2000), over 90% of the available datasets of non-electronic parts from RIAC are aggregated. In this document, two important variants of aggregate data, failure-censored and time-censored, are addressed and robust methods are proposed for data analysis purposes.

Reliability estimation based on complete failure-time data or failure-time data with certain types of censoring has been extensively studied in statistics and engineering. In practice, field data is convenient for a number of reasons such as happening in actual use conditions instead of laboratory and cost efficiency. When dealing with such data for reliability estimation, practitioners often face challenges of selecting the underlying failure-time distributions and the corresponding statistical inference methods. As mentioned, aggregate failure data usually exists in the forms of failure-censored aggregate data and time-censored aggregate data. Aggregate data has been addressed in the literature, nonetheless, only a few distributions are previously represented due to the complexities in statistical inference procedures. In this regards, presenting methods that are robust to the underlying distribution of failures are of great importance. In the rest of the document, failure-censored aggregate data is referred to as "aggregate data" and time-censored aggregate data is referred to as "censored aggregate data" for convenience.

The first chapter of this dissertation presents a new method for modeling lifetime data with covariates using phase-type (PH) distributions and a general life-stress relationship formulation. Lifetime data with covariates (e.g., temperature, humidity, and electric current) are frequently seen in engineering applications. An important example is accelerated life testing (ALT) data. In ALT, such data are collected by exposing test units of a product to severer-than-normal conditions to expedite product failure. The resulting lifetime and/or censoring data with covariates are often modeled by a probability distribution along with a life-stress relationship. However, if the prob-

ability distribution and the life-stress relationship selected cannot adequately describe the underlying failure process, the resulting reliability prediction will be misleading. A numerical study is presented to compare the performance of this method with a mixture of Weibull distributions model. This general method creates a new avenue to modeling and interpreting lifetime data with covariates for situations where the data-generating mechanisms are unknown or difficult to analyze using existing statistical tools.

In the second chapter PH distributions are utilized, for the first time, in reliability estimation based on aggregate failure-time data. Previously, only the Exponential, Normal, Gamma and Inverse Gaussian (IG) distributions had been used in analyzing aggregate failure-time data because these distributions have closed-form expressions for such data. However, the limited choices of probability distributions cannot satisfy extensive needs in a variety of engineering applications. Phase-type (PH) distributions are robust and flexible in modeling failure-time data as they can mimic a large collection of probability distributions of nonnegative random variables arbitrarily closely by adjusting the model structures. To this end, a maximum likelihood estimation (MLE) method and a Bayesian alternative are developed. For the MLE method, an expectation-maximization (EM) algorithm is developed to estimate the model parameters, and the corresponding Fisher information is used to construct the confidence intervals for the quantities of interest. For the Bayesian method, a procedure for performing point and interval estimation is also introduced. Several numerical examples show that the proposed PH-based reliability estimation methods are quite flexible and alleviate the burden of selecting a probability distribution when the underlying failure-time distribution is general or even unknown.

The third chapter presents a Phase-type distribution (PH) is for analyzing censored aggregate data for the first time. This type of data collection happens due to scheduled inspections during production. Censored aggregate data is equivalent to count data, where each component failure is equivalent to an event. First, a censored aggregate failure time model based on PH distribution is proposed. Then, an Expectation-Maximization (EM) algorithm for maximum likelihood (ML) parameter estimation and an alternative Bayesian RJMCMC method is developed. In many

situations count datasets are large and include covariates. Therefore, the model is presented and distinguished for cases without and with covariates. Numerical examples are provided for evaluation of the model and comparison with existing methods, showing the strength of the proposed method.

In the new era of data collection, reliability data from certain products are getting larger. As the technology and data transfer speed improves, more and more data is readily available on the product reliability from the customers. Important examples are electronic products where the manufacturer has direct access to the products' data. While this can significantly improve the reliability estimation and analysis, such data might be too large for the available computational capacities. Hence, there is a need for an efficient and proper data selection where the maximum possible information is gained from the data, while not processing it completely. In the fourth chapter, the problem of big data and data selection when using Phase-type (PH) distribution is investigated. In this chapter, a data selection method based on Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH). Fisher information is used as the information gain criteria. The performance of the model is evaluated using real-world data.

2 Robust Methods for Accelerated Life Testing Data Analysis

Lifetime data with covariates (e.g., temperature, humidity, and electric current) are frequently seen in engineering applications. An important example is accelerated life testing (ALT) data. In ALT, such data are collected by exposing test units of a product to severer-than-normal conditions to expedite product failure. The resulting lifetime and/or censoring data with covariates are often modeled by a probability distribution along with a life-stress relationship. However, if the probability distribution and the life-stress relationship selected cannot adequately describe the underlying failure process, the resulting reliability prediction will be misleading. This chapter develops a new method for modeling lifetime data with covariates using phase-type (PH) distributions and a general life-stress relationship formulation. A numerical study is presented to compare the performance of this method with a mixture of Weibull distributions model. This general method creates a new avenue to modeling and interpreting lifetime data with covariates for situations where the data-generating mechanisms are unknown or difficult to analyze using existing statistical tools.

2.1 Introduction

A covariate (e.g., temperature, humidity, and electric current) is a variable that is possibly predictive of the outcome under study. Data with covariates are frequently seen in engineering application. In particular, accelerated life testing (ALT) data are usually conducted by exposing test units of a product to severer-than-normal conditions to expedite the failure process. The resulting lifetime and/or censoring data are often modeled by a probability distribution along with a life-stress relationship. However, if the probability distribution and life-stress relationship selected cannot adequately describe the underlying failure process, the resulting reliability prediction will be misleading.

In practice, it is natural that there are underlying processes going through a series of stages before failures occur, and many of these processes are often partially or completely unobservable due to technology barriers or lack of understanding of failure mechanisms (Kuo, 2006). Despite

some model-selection guidelines (Sethuraman & Singpurwalla, 1982), choosing adequate distributions to fit such data is always a challenging task. Although some commercial software packages provide several options for probability distributions or even a distribution-selection wizard based on the likelihood values of a limited number of candidate models, no generic model-construction and -selection methods have been reported in the related literature. To seek new mathematical and statistical tools to facilitate the modeling of such data, a critical question to be asked is: Can we find a family of versatile probability distributions along with a general life-stress relationship to model complex lifetime data with covariates?

A continuous phase-type (PH) distribution describes the time to absorption of a continuous-time Markov chain (CTMC) defined on a finite-state space. Since the class of PH distributions is dense, any distribution defined on $[0, \infty)$, in principle, can be approximated arbitrarily closely by a PH distribution (Asmussen et al., 1996; Johnson & Taaffe, 1990). In reliability engineering, Ruiz-Castro et al., 2008 investigated a repairable cold-standby system using a quasi-birth-and-death process. Jonsson et al., 1994 used PH distributions to deal with non-exponential lifetime distributions in a system. Recently, to model ALT data, Liao and Guo, 2013 explored a new accelerated failure time model (Bagdonavicius & Nikulin, 2001) based on Erlang-Coxian distributions (Osogami & Harchol-Balter, 2003) to characterize ALT data. The ALT model belongs to an accelerated failure time (AFT) model, which incorporates the effect of a covariate on the product's reliability through time scaling.

In this chapter, a more general method is proposed for modeling lifetime data with covariates. In this method, a general life-stress relationship is introduced into the Coxian distribution and a maximum likelihood-based approach is utilized to estimate the model parameters and perform model selection. A comparison study is conducted to illustrate the modeling capability of this method. The unique contribution of this work is that it provides a flexible method for modeling and interpreting lifetime data with covariates when the underlying failure mechanisms are unknown or difficult to analyze using the traditional statistical tools.

The remainder of this chapter is organized as follows. Section 2.2 describes the proposed

model based on the Coxian distribution and the widely used model based on a mixture of Weibull distributions. The corresponding model selection aspects are provided on Section 2.3. In Section 2.4, a numerical example is provided to compare the performance of the proposed method and the mixture of Weibull distributions model. Finally, conclusions are drawn in Section 2.5.

2.2 General Methods for Modeling Lifetime Data with Covariates

2.2.1 Use of a Mixture of Weibull Distributions

A mixture of Weibull distributions is widely used in modeling complex lifetime data that may not be well described by a single probability distribution such as the two- or three-parameter Weibull distribution, lognormal distribution, and Gamma distribution. Moreover, it has also been used to approximate other probability distributions.

The probability density function (PDF) $f(t;\underline{\theta})$ of a mixture of m Weibull distributions can be expressed as:

$$f(t;\underline{\theta}) = \sum_{i=1}^{m} p_i \frac{\beta_i}{\eta_i}\left(\frac{t}{\eta_i}\right)^{\beta_i - 1} \exp\left(-\left(\frac{t}{\eta_i}\right)^{\beta_i}\right), \tag{2.1}$$

where $\sum_{i=1}^{m} p_i = 1$, β_i and η_i are the shape and scale parameters of the individual Weibull distribution, and $\underline{\theta}$ represent a vector containing all model parameters. The corresponding cumulative distribution function (CDF) $F(t;\underline{\theta})$ is:

$$F(t;\underline{\theta}) = 1 - \sum_{i=1}^{m} p_i \exp\left(-\left(\frac{t}{\eta_i}\right)^{\beta_i}\right). \tag{2.2}$$

To quantify the effects of a covariate z on the model parameters, a general approach is to model

each individual parameter as a function of the covariate as:

$$F(t;\underline{\tilde{\theta}},z)=1-\sum_{i=1}^{m}p_i\exp\left(-\left(\frac{t}{\eta_i(z;\underline{\alpha_i})}\right)^{\beta_i(z;\underline{\gamma_i})}\right). \tag{2.3}$$

For example, a widely used life-stress relationship takes a log-linear form $\eta_i(z;\underline{\alpha_i})=\exp(\alpha_{0i}+\alpha_{1i}z)$ and assumes a constant shape parameter $\beta_i(z;\underline{\gamma_i})=\beta_i$.

2.2.2 Use of Coxian Distributions

We consider a CTMC with finite states $1,2,\ldots,N,N+1$, where state $N+1$ is the only absorbing state and the others are transient states. Let T be the time to absorption of the CTMC. The Coxian distribution is a versatile mathematical model that describes the absorbing time of a CTMC. Fig. 2.1 shows a three-phase Coxian model.

The infinitesimal generator of is given by:

$$\boldsymbol{Q}=\begin{bmatrix}\boldsymbol{S} & \boldsymbol{q}\\ \boldsymbol{0} & 0\end{bmatrix} \tag{2.4}$$

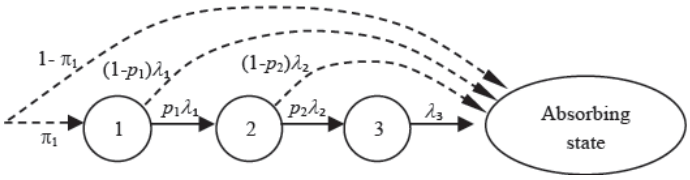

Figure 2.1: CTMC for a three-phase Coxian distribution.

For an acyclic PH distribution (e.g., Coxian), $\boldsymbol{S}$ is an upper triangular matrix, $\boldsymbol{q}=-\boldsymbol{S}\boldsymbol{e}$, $\boldsymbol{\pi}=[1,0,\cdots,0]$, and $\boldsymbol{e}=[1,1,\cdots,1]'$. Length of $\boldsymbol{\pi}$ is N. For the N-phase Coxian distribution,

we have:

$$\boldsymbol{S} = \begin{bmatrix} -\lambda_1 & \lambda_1 p_1 & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 p_2 & 0 \\ \vdots & 0 & \ddots & \vdots \\ 0 & \cdots & 0 & -\lambda_N \end{bmatrix}; \boldsymbol{q} = \begin{bmatrix} \lambda_1(1-p_1) \\ \lambda_2(1-p_2) \\ \vdots \\ -\lambda_N \end{bmatrix}. \tag{2.5}$$

Let the process $X(t)_{t>=0}$ start in the first phase by making $\pi_1 = 1$. Then, the CDF and PDF of the time to absorption T can be expressed as: $F(t) = 1 - \boldsymbol{\pi} exp(t\boldsymbol{S})\boldsymbol{e}$ and $f(t) = \boldsymbol{\pi} exp(t\boldsymbol{S})\boldsymbol{q}$, respectively, where $exp(.)$ represents matrix exponential. A general life-stress relationship can be introduced as follows:

$$\lambda_i(z) = \Psi_i(z; \underline{\gamma}_i) \tag{2.6}$$

where each transition rate λ_i is modeled as a function of the covariate with parameter $\underline{\gamma}_i$. Because all λ_i, $i = 1, 2, \cdots, N$, are positive, a useful candidate is $\Psi_i = exp(\gamma_{0i} + \gamma_{1i} z)$. In this chapter, this exponential form of life-stress relationship is adopted.

2.3 Model Selection and Parameter Estimation

2.3.1 Maximum Likeliood Estimation Method

To estimate the model parameters $\underline{\tilde{\theta}}$ for a mixture of m Weibull distributions, the maximum likelihood estimation method can be used. Given a set of lifetime data collected under J levels of the covariate $\{t_{jk}, \delta_{jk}, z_j\}$, where $\delta_{jk} = \{1, \text{if } t_{jk} \text{ is lifetime}; 0, \text{otherwise}\}$, the likelihood function can be expressed as:

$$l(\underline{\tilde{\theta}}) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} \left[\sum_{i=1}^{m} \begin{array}{l} p_i \frac{\beta_i(z_j;\underline{\gamma}_i)}{\eta_i(z_j;\underline{\alpha}_i)} \left(\frac{t_{jk}}{\eta_i(z;\underline{\alpha}_i)} \right)^{\beta_i(z_j;\underline{\gamma}_i)-1} \\ \times \exp\left(-\left(\frac{t_{jk}}{\eta_i(z_j;\underline{\alpha}_i)} \right)^{\beta_i(z_j;\underline{\gamma}_i)} \right) \end{array} \right]^{\delta_{jk}} \times \left[\sum_{i=1}^{m} p_i \exp\left(-\left(\frac{t_{jk}}{\eta_i(z_j;\underline{\alpha}_i)} \right)^{\beta_i(z_j;\underline{\gamma}_i)} \right) \right]^{1-\delta_{jk}}. \tag{2.7}$$

where K_j is the total number of observations under the jth level of covariate. Then, the maximum likelihood estimate of $\underline{\tilde{\theta}}$ can be obtained by maximizing the log-likelihood function $ln\, l(\underline{\tilde{\theta}})$ after taking the natural-log of equation 2.7.

Similarly, to estimate the model parameters in a specific Coxian-based model, the likelihood function is given by:

$$l(\underline{\tilde{\theta}}) = \prod_{j=1}^{J} \prod_{k=1}^{K_j} [\boldsymbol{\pi} \exp(t_{jk}\boldsymbol{S}(z_j))\boldsymbol{q}(z_j)]^{\delta_{jk}} [\boldsymbol{\pi} \exp(t_{jk}\boldsymbol{S}(z_j))\boldsymbol{e}]^{1-\delta_{jk}}, \tag{2.8}$$

where matrices $\boldsymbol{S}(z_j)$ and $\boldsymbol{q}(z_j)$ have the forms given in equation 2.5 with all $\lambda_i(z), i = 1, 2, \cdots, N$, as described by equation 2.6. In this research, an expectation-maximization algorithm is utilized to maximize the corresponding log-likelihood function for obtaining the maximum likelihood estimates of the model parameters.

2.3.2 Likelihood-based Model Selection

It is worth pointing out that when either the Coxian-based model or the mixture of Weibull distributions is used to model lifetime data with covariates, the model structures are not necessarily pre-determined. Instead, a model selection method can be utilized to determine the models that provide the best fit to the data. In particular, for the Coxian-based model the number of phases can be determined according to the change in the value of maximum likelihood as the number of phases is increased. Similarly, for the mixture of Weibull distributions model the number of Weibull distributions in the model can be obtained by investigating the values of maximum likelihood as more complex models are considered.

An alternative for such model selection problems under the maximum likelihood estimation framework is the use of Akaike Information Criterion (AIC):

$$AIC = 2p - 2\ln l(\underline{\hat{\theta}}) \tag{2.9}$$

where p is the number of parameters in parameter vector $\underline{\theta}$ of the corresponding model and $\underline{\hat{\theta}}$ is

the maximum likelihood estimate of $\underline{\theta}$. The common practice is to choose the model that has the lowest AIC value compared to other candidate models.

2.4 Numerical Example

In this section, we use the ALT data reported by Liao and Elsayed, 2010 to illustrate the use of the proposed method for modeling lifetime data with a single covariate.

2.4.1 Experimental Setup

The purpose of this ALT experiment is to estimate the reliability of a type of miniature lamps under the use condition: 2 volts. The highest operating voltage of the lamp is 5 volts. Three constant voltage levels were used in the experiment: 5 volts, 3.5 volts, and 2 volts. After standardization $((volts - 2)/3)$, the three voltage levels are: $z_1 = 1$, $z_2 = 0.5$, and $z_3 = 0$. The observed lifetimes and censoring times are presented in Table 2.1. To avoid making an assumption on the underlying distribution, the proposed Coxian-based model as well as the mixture of Weibull distributions model are utilized to predict the reliability of this type of miniature lamps.

2.4.2 Results of Mixture of Weibull Distributions

Figures 2.2 and 2.3 show the estimated CDFs using the Kaplan-Meier estimator, a mixture model of three Weibull distributions, and a mixture model of four Weibull distributions. Based on equation 2.5, the corresponding numbers of parameters of the two mixture models are fourteen and nineteen, respectively. However, by comparing the result in the work of Liao and Elsayed, 2010 where the lognormal distribution is utilized, the performance of the mixture model is inferior. On the other hand, from Figure 2.4 that illustrates the shape and scale parameters in the mixture model of four Weibull distributions, the adopted life-stress relationships appear to be appropriate.

2.4.3 Results of the Proposed Coxian-based General Method

To demonstrate the superior performance of the proposed Coxian-base method with the general life-stress relationship formulation (exponential functions), models with different numbers of phases are obtained. Figures 2.5 and 2.6 illustrate the estimation performance of the three-phase Coxian-based model and the seven-phase alternative. From equation 2.5 and 2.6, the numbers of model parameters of the two models are eight and twenty, respectively. One can see that the proposed method is able to provide quite adequate statistical fits. From Figure 2.7, the three-phase Coxian-based model is suggested based on the AICs.

Table 2.1: ALT Data of a Type of Miniature Lamp.

Stress	Lifetime in hours ("+" censored)							
5V ($z_1 = 1$)	20.5	22.3	23.2	24.7	26	34.1	39.6	41.8
	43.6	44.9	47.7	61.6	62.1	65.5	70.8	87.8
	118.3	120.1	145.4	157.4	180.9	187.7	204	206.7
	213.9	215.2	218.7	254.1	262.6	293	304	313.7
	314.1	317.9	337.7	430.2				
3.5V ($z_2 = 0.5$)	37.8	43.6	51.1	58.6	65.5	65.9	75.6	82.5
	88.1	89	106.6	113.1	121.1	121.5	128.3	151.8
	171.7	181	202.7	211.7	230.7	249.9	275.6	285
	296.2	358.5	379.8	434.5	493.1	506.1	570	577.7
	876.3	890+	890+	922	941+	941+		
2V ($z_3 = 0$)	223.1	254	316.7	560.2	679	737	894.4	930.5+
	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+
	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+

2.5 Conclusion

This research addresses general methods for modeling lifetime data with covariates. A Coxian-based method that incorporates a flexible approach to modeling life-stress relationship is proposed and compared with a widely used alternative based on a mixture of Weibull distributions. The advantage of the proposed general data analysis method is that an adequate fit to lifetime data can be obtained by gradually changing the number of phases of the associated CTMC. Without assuming

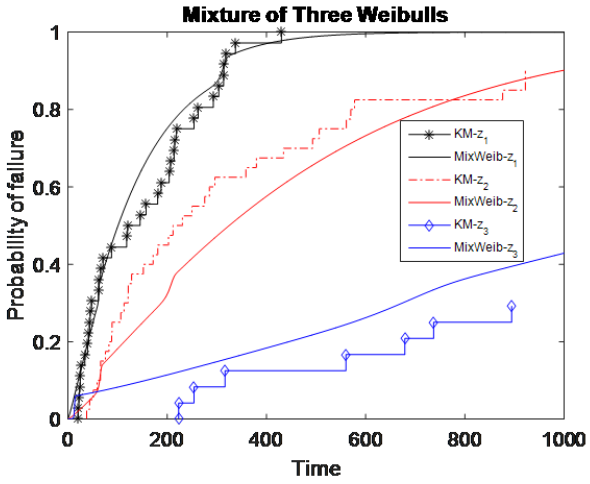

Figure 2.2: Statistical fits of the mixture model of three Weibull distributions.

other particular probability distributions for the lifetimes, such as extreme-value distributions, log-normal distribution, and gamma distributions, this method leads to a Coxian-based model which can well represent the underlying lifetime distribution that may be difficult to model. To automatically determine the model structure, a maximum likelihood-based approach is developed for adaptively determining the number of phases and estimating the model parameters. The numerical example demonstrates that the proposed general method indeed provides practitioners with a convenient statistical tool for modeling lifetime data with covariates. Compared with the mixture of Weibull distributions model, the proposed method is able to provide more accurate estimates with comparable model complexity.

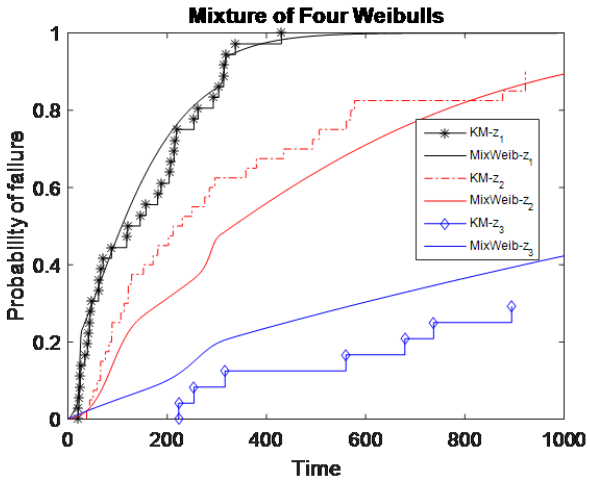

Figure 2.3: Statistical fits of the mixture model of four Weibull distributions.

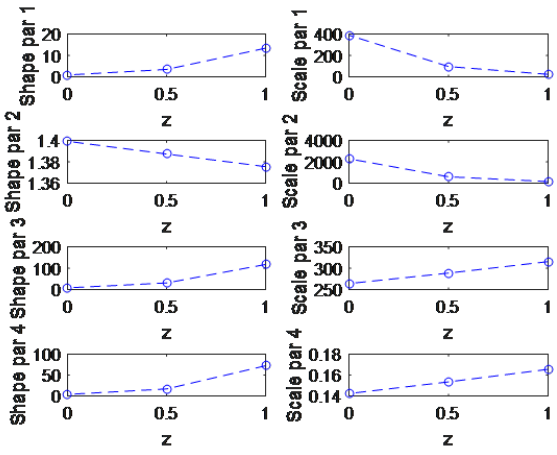

Figure 2.4: Shape and scale parameters in a mixture of four Weibull distributions.

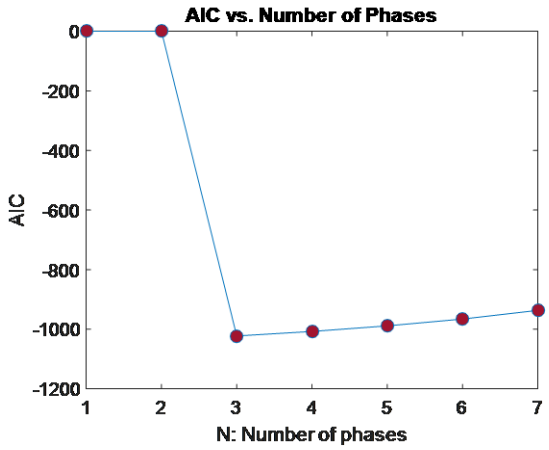

Figure 2.5: AIC values for Coxian-based models with different numbers of phases.

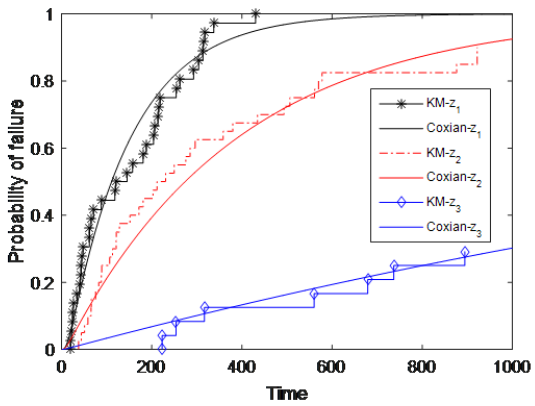

Figure 2.6: Statistical fits of the three-phase Coxian model with general life-stress relationship.

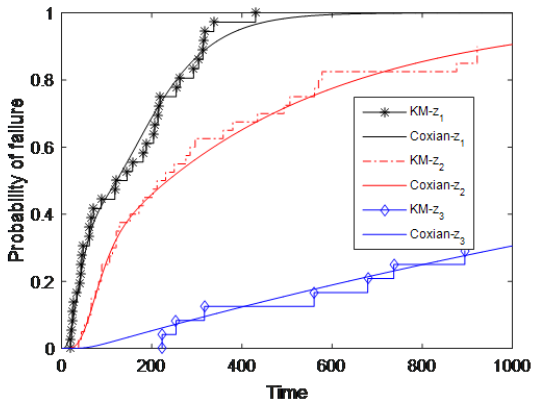

Figure 2.7: Statistical fits of the seven-phase Coxian model with general life-stress relationship.

References

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, *23*(4), 419–441.

Bagdonavicius, V., & Nikulin, M. (2001). *Accelerated life models: modeling and statistical analysis*. CRC press.

Johnson, M. A., & Taaffe, M. R. (1990). Matching moments to phase distributions: Density function shapes. *Stochastic Models*, *6*(2), 283–306.

Jonsson, E., Andersson, M., & Asmussen, S. (1994). A practical dependability measure for degradable computer systems with non-exponential degradation. *IFAC Proceedings Volumes*, *27*(5), 227–233.

Kuo, W. (2006). Challenges related to reliability in nano electronics. *IEEE Transactions on Reliability*, *55*(4), 569–570.

Liao, H., & Elsayed, E. A. (2010). Equivalent accelerated life testing plans for log-location-scale distributions. *Naval Research Logistics (NRL)*, *57*(5), 472–488.

Liao, H., & Guo, H. (2013). A generic method for modeling accelerated life testing data. *2013 Proceedings Annual Reliability and Maintainability Symposium (RAMS)*, 1–6.

Osogami, T., & Harchol-Balter, M. (2003). A closed-form solution for mapping general distributions to minimal PH distributions. *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, 200–217.

Ruiz-Castro, J. E., Pérez-Ocón, R., & Fernández-Villodre, G. (2008). Modelling a reliability system governed by discrete phase-type distributions. *Reliability Engineering & System Safety*, *93*(11), 1650–1657.

Sethuraman, J., & Singpurwalla, N. D. (1982). Testing of hypotheses for distributions in accelerated life tests. *Journal of the American Statistical Association*, *77*(377), 204–208.

3 Flexible Methods for Reliability Estimation Using Aggregate Failure-time Data

Reliability estimation based on complete failure-time data or failure-time data with certain types of censoring has been extensively studied in statistics and engineering. However, the actual failure times of individual components are usually unavailable in many applications. Instead, only aggregate failure-time data are collected by actual users due to technical and/or economic reasons. When dealing with such data for reliability estimation, practitioners often face challenges of selecting the underlying failure-time distributions and the corresponding statistical inference methods. So far, only the Exponential, Normal, Gamma and Inverse Gaussian (IG) distributions have been used in analyzing aggregate failure-time data because these distributions have closed-form expressions for such data. However, the limited choices of probability distributions cannot satisfy extensive needs in a variety of engineering applications. Phase-type (PH) distributions are robust and flexible in modeling failure-time data as they can mimic a large collection of probability distributions of non-negative random variables arbitrarily closely by adjusting the model structures. In this chapter, PH distributions are utilized, for the first time, in reliability estimation based on aggregate failure-time data. To this end, a maximum likelihood estimation (MLE) method and a Bayesian alternative are developed. For the MLE method, an expectation-maximization (EM) algorithm is developed to estimate the model parameters, and the corresponding Fisher information is used to construct the confidence intervals for the quantities of interest. For the Bayesian method, a procedure for performing point and interval estimation is also introduced. Several numerical examples show that the proposed PH-based reliability estimation methods are quite flexible and alleviate the burden of selecting a probability distribution when the underlying failure-time distribution is general or even unknown.

3.1 Introduction

3.1.1 Background

The accuracy and authenticity of failure-time data play an important role in the reliability analysis of a product. One source of failure-time data is from laboratory life tests. However, a common issue in using laboratory test data is that some of unknown influential factors (e.g., ambient temperature, humidity, corrosive gas, ultraviolet light) exposed by the product in the field may be ignored in the tests. As a result, the outcome of laboratory tests may not be consistent with the behavior of the product's lifetime in the field. Another source of failure-time data is provided by the actual users of the product. Obviously, it is quite valuable to utilize the rich sources of field data for product reliability estimation as such data reflect the behavior of the product under the real usage conditions. For this reason, organizations, such as the U.S. Department of Defense, have collected a large volume of failure data (OREDA, 2009; Mahar et al., 2011; Denson et al., 2014).

One hurdle of using field data is that the exact failure times of individual components are usually unavailable. In many applications, the collected data contains the number of failed components in a single position of a system along with the system's cumulative operating time until the last failure. This type of data is called aggregate failure-time data. Figure 3.1 shows an example of aggregate data. For a specific component in each system, the user replaces it whenever it fails without recording the actual failure time. Eventually, a data point representing the time from the first installation to the last component failure (e.g., m_2 failures [replacements] in System #2) is reported. Compared to laboratory testing data with actual failure times, the aggregate data is more concise (Chen and Ye, 2017). To estimate the product reliability from such aggregate data, only a few probability distributions (i.e., Exponential, Normal, Gamma and Inverse Gaussian (IG)) have been used because their closed-form expressions for aggregate data are available. For other widely used probability distributions, such as Weibull and Lognormal, the closed-form expressions are not attainable. Apparently, the limited choices of probability distributions cannot satisfy extensive needs in many engineering applications where only aggregate data are reported. To assist prac-

titioners in using abundant aggregate data, it is necessary to develop a flexible approach and the corresponding statistical inference methods beyond the use of limited probability distributions for reliability estimation.

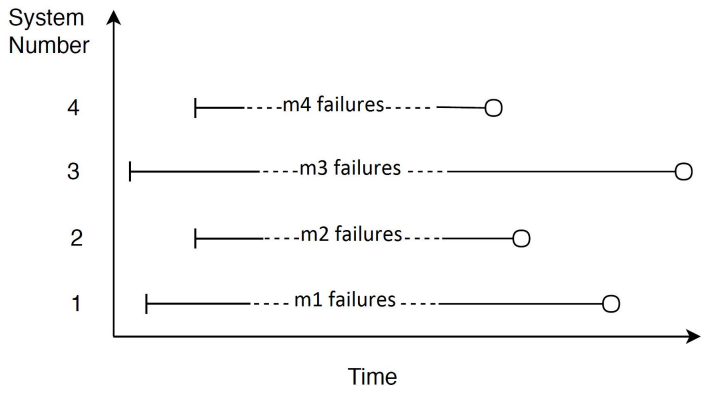

Figure 3.1: An example of aggregate data: the number of failures (e.g., m_1, ..., m_4) and the time interval from the first installation till the death of the last component.

Phase-type (PH) distributions are robust and flexible in modeling failure-time data as they can mimic a large collection of probability distributions of nonnegative random variables arbitrarily closely by adjusting the model structures. In this chapter, PH distributions are utilized to model aggregate data for the first time. A new expectation-maximization (EM) algorithm is developed to obtain the maximum likelihood estimates (MLE) of model parameters. A Bayesian alternative is also introduced to incorporate prior knowledge in parameter estimation. For both methods, the interval estimates for the quantities of interest are derived.

3.1.2 Related Work

The Exponential distribution has been widely used in reliability for modeling failure-time data. Because of its tractability, aggregate data are often collected and analyzed using this distribution. Coit and Dey (1999) developed an approach to analyze Type II censored data when individual failure times were not available. They also presented a hypothesis test to examine the Exponential distribution assumption and tested their specific data set from a Weibull distribution. Regarding the use of Gamma distribution, Coit and Jin (2000) developed an MLE procedure for handling ag-

gregate failure-time data. A quasi-Newton method was used to find the MLE of model parameters. Chen and Ye (2017a) proposed random effects models based on the Gamma and IG distributions to handle aggregate data. Later, Chen and Ye (2017b) provided a collection of approaches to handle aggregate data using the Gamma and IG distributions. It is worth pointing out that interval estimation of quantity of interest using individual failure-time data has been extensively studied (see Bhaumik et al., 2009), but much less effort has been taken on the analysis of aggregate data. Chen and Ye (2017b) proposed powerful interval estimation algorithms for the Gamma and IG distributions using aggregate data. An extension to the analysis of aggregate lifetime data is the modeling of time-censored aggregate data. This type of data is also abundant, and Chen et al. (2019) proposed models for the analysis of this type of data under a Bayesian framework. Moreover, Chen et al. (2020) proposed a parametric framework for analysis time-censored aggregate data under Gamma and IG distributions.

When dealing with aggregate data using probability distributions other than the Exponential, Normal, Gamma and IG distributions, an intuitive idea is to perform distribution approximation. To approximate probability distributions for data analysis, extensive studies have been focused mainly on the use of Lognormal distribution (Beaulieu and Rajwani, 2004; Beaulieu and Xie, 2004; Lam and Le-Ngoc, 2007; Mehta et al., 2007; Cobb et al.; 2012, Asmussen et al., 2016), mixture of Weibull distributions (Bučar et al., 2004; Jin and Gonigunta, 2010; Elmahdy and Aboutahoun, 2013) and the Laplace method (Rizopoulos et al.; 2009, Rue et al., 2009; Asmussen et al., 2016). Moreover, PH distributions are proved to be able to approximate a large collection of probability distributions of non-negative random variables arbitrarily closely. Because of this, a large amount of work has been done on approximating general distributions with PH distributions. Phase-type probability approximation was basically done by matching the first two moments of a PH distribution with those of a target distribution. This can be done by using a two-phase hyper-exponential distribution for approximating the distributions with square coefficient of variation (C^2) greater than 1 and Erlang distribution for those with C^2 less than 1. Another possible moment-matching method is using the Coxian distribution for distributions with $C^2 > 1$ and using the generalized

Erlang distribution for those with $C^2 < 1$. Also, solutions for matching three moments only for distributions with $C^2 > 1$ through a two-phase hyper-exponential distribution was used. Two-phase Coxian and mixed Erlang distributions were also applied for the same purpose. Phase-type distribution approximation started to significantly improve in terms of generality when Telek and Heindl (2003) matched two-phase acyclic PH distributions with no mass probability at 0 for distributions with $C^2 \geq \frac{1}{2}$. To make approximation more accurate and general, Osogami and Harchol-Balter (2003) and Osogami and Harchol-Balter (2006) proposed an algorithm for mapping a general distribution to a PH distribution by matching the first three moments. Horvath and Telek (2007) proposed an approximation approach for matching the first $2N - 1$ moments for an acyclic PH distribution with N phases. Other than moments matching, some studies have been focused on matching the shape of a desired distribution via PH approximation (Starobinski and Sidi (2000), Riska et al. (2004)).

PH distributions have been applied in queueing, healthcare, risk analysis, and reliability. In the area of reliability, Delia and Rafael (2008) modeled a deteriorating system involving both internal and external failures and applied PH distributions to two different repair types. Kharoufeh et al. (2010) introduced a hybrid, degradation-based component reliability model considering environmental effects by PH distribution. Segovia and Labeau (2013) investigated the reliability of a multi-state system subject to internal wear-out and external shocks using a PH distribution. Liao and Guo (2013) modeled accelerate life testing (ALT) data using the Erlang-Coxian distribution. Liao and Karimi (2017) studied a more flexible method of analyzing ALT data using a PH distribution. In the literature, however, PH distributions have never been utilized in modeling aggregate failure-time data. To alleviate the burden of selecting probability distributions and provide a flexible means for data analysis, this chapter studies the use of PH distributions in modeling aggregate failure-time data for the first time. Recently, researchers have used PH distribution for modeling multi-state systems as well as degradation analysis. Cui and Wu (2019) used PH distribution for modeling Marcov repairable systems. Li et al. (2019) and Xu et al. (2020) studied deteriorating structures under aging or shocks using PH distributions.

A technical challenge of using PH distributions is model parameter estimation. Asmussen et al. (1996) developed an EM algorithm to obtain the MLE of model parameters. They also used the EM algorithm to minimize information divergence in density approximation. Since the EM algorithm is computationally intensive, Okamura et al. (2011) proposed a refined EM algorithm to reduce the computational time using uniformization and an improved forward-backward algorithm. As an alternative, under the framework of Bayesian statistics, Bladt et al. (2003) used a Markov chain Monte Carlo (MCMC) method combined with Gibbs sampling for general PH distributions. Watanabe et al. (2012) also presented an MCMC approach to fit PH distributions while using uniformization and backward likelihood computation to reduce the computational time. Ausín et al. (2008) and McGrory et al. (2009) explored two special cases of PH distributions (i.e., Erlang and Coxian) through a Reversible Jump Markov chain Monte Carlo (RJMCMC) method. Yamaguchi et al. (2010), and Okamura et al. (2014) presented variational Bayesian methods to improve the computational efficiency of PH estimation in comparison to MCMC. It is worth pointing out that all of these estimation methods were not developed for aggregate data. Perreault et al. (2015) proposed a swarm-based approach for learning PH distributions for continuous time Bayesian networks. In this chapter, efforts will be focused on developing a collection of new MLE and Bayesian methods for the analysis of aggregate failure-time data.

3.1.3 Overview

The remainder of this chapter is organized as follows. Section 3.2 introduces PH distributions. Section 3.3 provides the statistical procedures of the proposed MLE method, including the EM algorithm and the use of Fisher information for interval estimation. The Bayesian alternative is presented in Section 3.4 for both parameter and credible interval estimation. In Section 3.5, numerical examples are provided to illustrate the practical use of the proposed PH-based aggregate data analysis methods. A simulation study shows the strength of PH distribution in dealing with aggregate data from an arbitrarily selected probability distribution, and the coverage probability of Normal approximate interval estimation method is compared with the one obtained via nonpara-

metric bootstrapping. In addition, a real dataset is also analyzed to demonstrate the practical use of the proposed methods in industrial statistics. Finally, conclusions are drawn in Section 3.6.

3.2 PH Distributions

A PH distribution describes the time to absorption of a Continuous-time Markov Chain (CTMC) defined on a finite-state space. Consider a finite-state CTMC $X(t)_{t\geq 0}^{\infty}$ with N transient states and an absorbing state $N+1$, then the CTMC with the specific structure can be described by an infinitesimal generator matrix:

$$\mathbb{Q} = \begin{pmatrix} 0 & \mathbf{0}' \\ \mathbf{S}^0 & \mathbb{S} \end{pmatrix}. \tag{3.1}$$

where $\mathbf{0}' = [0, ..., 0]$, $\mathbb{S}$ is the subgenerator matrix of the transition rates between the transient states, and $\mathbf{S}^0 = -\mathbb{S}\mathbf{1}$ represents the absorption rates with $\mathbf{1} = [1, ..., 1]^T$ (Buchholz et al., 2014). In particular, the transition rate matrix of an acyclic CTMC can be expressed as:

$$\mathbb{S} = \begin{pmatrix} -\lambda_1 & p_{12}\lambda_1 & p_{13}\lambda_1 & \cdots & p_{1N}\lambda_1 \\ 0 & -\lambda_2 & p_{23}\lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -\lambda_{N-1} & p_{(N-1)N}\lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\lambda_N \end{pmatrix}, \tag{3.2}$$

where $0 \leq p_{ij} \leq 1$, $i < j$, $i = 1, 2, ..., N-1$, $j = 1, 2, ..., N$, and $\sum_{j=1}^{N} p_{ij} \leq 1$.

Figure 3.2 shows the corresponding CTMC of the general Phase type distribution. The probability density function (PDF) and cumulative distribution function (CDF) of PH distribution are:

$$f(t) = \boldsymbol{\pi} e^{\mathbb{S}t}\mathbf{S}^0, \qquad F(t) = 1 - \boldsymbol{\pi} e^{\mathbb{S}t}\mathbf{1}, \tag{3.3}$$

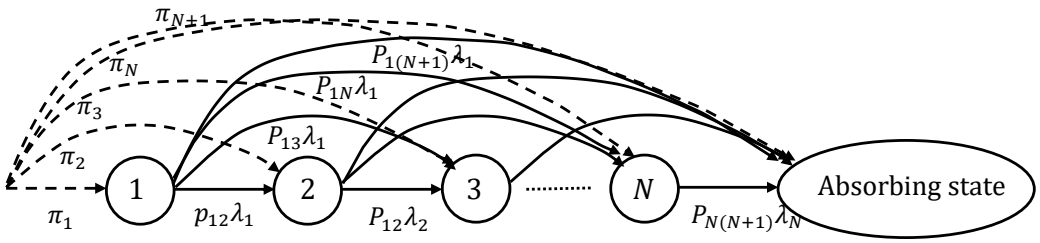

Figure 3.2: CTMC of an N-Phase Phase type distribution.

respectively, where $\boldsymbol{\pi} = [\pi_1, ..., \pi_k, ..., \pi_N]$ is the initial probability vector with $\sum_{k=1}^{N} \pi_k = 1$ that describes the probability of the process being started in each phase.

The most popular PH distributions are the Exponential, Erlang, Hyper-exponential, Hypo-exponential, Hyper-Erlang, and Coxian distributions. Specially, Coxian distribution has been widely used for resolving the non-identifiability problem of PH distributions. Figure 3.3 shows the CTMC of an N-phase Coxian distribution. The transition rate matrix of an N-phase Coxian distribution is sparse, which has zero p_{ij}'s, except $p_{i(i+1)}$'s for $i = 1, 2, \cdots, N-1$.

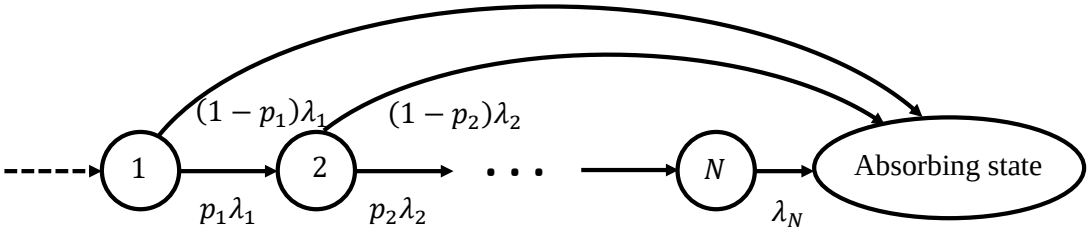

Figure 3.3: CTMC of an N-Phase Coxian distribution.

In this work, for the purpose of parameter estimation, the following reparameterization is used for the transition rate matrix of an N-phase Coxian distribution. In this reparameterization $\lambda_i \equiv$

$p_i\lambda_i$ and $\mu_i \equiv (1-p_i)\lambda_i$.

$$\mathbb{S} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -(\lambda_{N-1}+\mu_{N-1}) & \lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\mu_N \end{pmatrix}. \tag{3.4}$$

Then, the absorption rate matrix can be expressed as $\mathbf{S}^0 = [\mu_1, \mu_2, ..., \mu_N]^T$. In practice, Coxian distribution emerges as a very flexible distribution while carrying considerably less parameters than general PH distributions. Indeed, the number of parameters in a general PH distribution is $O(N^2)$ while for Coxian, it becomes $O(N)$ which justifies the use of Coxian in practice. Moreover, it can be shown that any acyclic PH distribution can be converted to a Coxian distribution. Because of its flexibility and structural simplicity, Coxian distribution is used in this chapter although the proposed methods can be applied to other PH distributions.

3.3 Maximum Likelihood Estimation

3.3.1 EM Algorithm for Individual Failure-time Data

An EM algorithm for estimating the parameters of a PH distribution was first proposed by Asmussen et al. (1996). Given each individual failure time, one needs to deal with having a number of unobserved sojourning times in those transient states of CTMC. The likelihood function can be rewritten as:

$$\mathcal{L}((\boldsymbol{\pi}, \mathbb{S})|\boldsymbol{\tau}) = f(z|(\boldsymbol{\pi}, \mathbb{S})) = \prod_{i=1}^{N} \pi(i)^{B_i} \prod_{i=1}^{N} e^{Z_i \mathbb{S}(i,i)} \prod_{i=1}^{N} \prod_{j=1}^{N+1} \mathbb{S}(i,j)^{N_{ij}}, \tag{3.5}$$

where $\boldsymbol{\tau} = (t_1, t_2, \cdots, t_M)$ contains M observed individual failure times, z represents the complete observation, and B_i, N_{ij} and Z_i are the missing values of the data representing the number of times the Markov process started in phase i, the number of jumps from phase i to phase j, and the

total time spent in phase i, respectively, for $i = 1, 2, ..., N$ in an N-phase PH distribution.

Note that this likelihood function is evaluated using the estimated values of the unobserved data obtained in the Expectation step (E-step). To do this, a few statistics are defined in advance:

$$\mathbf{f}_{(\boldsymbol{\pi},\mathbb{S}),t} = \boldsymbol{\pi} e^{\mathbb{S}t}, \qquad \mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t} = e^{\mathbb{S}t}\mathbf{S}^0, \qquad \mathbf{F}_{(\boldsymbol{\pi},\mathbb{S}),t} = \int_0^t (\mathbf{f}_{(\boldsymbol{\pi},\mathbb{S}),t-u})^T (\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),u})^T du. \tag{3.6}$$

Then, the conditional expectation of unobserved variables are calculated using the current estimates of model parameters as:

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[B_i] = \frac{1}{M}\sum_{k=1}^{M} \frac{\boldsymbol{\pi}(i)\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}(i)}{\boldsymbol{\pi}\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}}, \tag{3.7}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i] = \frac{1}{M}\sum_{k=1}^{M} \frac{\mathbf{F}_{(\boldsymbol{\pi},\mathbb{S}),t_k}(i,i)}{\boldsymbol{\pi}\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}}, \tag{3.8}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{ij}] = \frac{1}{M}\sum_{k=1}^{M} \frac{\mathbb{S}(i,j)\mathbf{F}_{(\boldsymbol{\pi},\mathbb{S}),t_k}(i,j)}{\boldsymbol{\pi}\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}}, \tag{3.9}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{in+1}] = \frac{1}{M}\sum_{k=1}^{M} \frac{\mathbf{S}^0(i)\mathbf{f}_{(\boldsymbol{\pi},\mathbb{S}),t_k}(i)}{\boldsymbol{\pi}\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}}, \tag{3.10}$$

where $i, j = 1, 2, \cdots, N$. In the M-step, the parameters of the distribution are re-estimated using the current estimate of the complete data (Buchholz et al., 2014):

$$\hat{\boldsymbol{\pi}}(i) = E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[B_i], \qquad \hat{\mathbb{S}}(i,j) = \frac{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{ij}]}{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i]},$$
$$\hat{\mathbf{S}}^0(i) = \frac{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{in+1}]}{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i]}, \qquad \hat{\mathbb{S}}(i,i) = -(\hat{\mathbf{S}}^0(i) + \sum_{i\neq j}^{n} \hat{\mathbb{S}}(i,j)). \tag{3.11}$$

Note that this EM algorithm monotonically improves the likelihood value to achieve the MLE of model parameters. However, it was developed only for individual failure-time data.

3.3.2 The Proposed EM Algorithm for Aggregate Data

The previous EM algorithm uses each data point in the E-step to contribute to estimating the unobserved or missing values. As it can be seen, from each data point one value for each Z_i, B_i

and N_{ij} can be found "for each phase" of the distribution, and the mean values of these give the expected values of the variables in the E-step.

The challenge of using aggregate data, however, is that each data point corresponds to PH distributions with different numbers of failures. This causes the underlying distributions for different data points to have different numbers of phases. Unlike individual failure-time data, in this case we have independent but not identically distributed variables. Considering m_k as the number of failed components for data point k, the data point follows a PH distribution with Nm_k phases. As a result, the transition rate matrix for m_k failures is an $(Nm_k) \times (Nm_k)$ matrix. So, it is necessary to determine how and for which phases those variables should be estimated (Karimi et al. (2019)).

Primarily, the most important aspects are finding the resulting transition rate matrix for the sum of a number of similar N-phase PH variables and deriving the properties of the resulting distribution to develop an EM algorithm for the case of aggregate data. For some distributions, such as Gamma, this is straightforward. In the case of sum of m similar Gamma variables, the resulting variable will follow a Gamma distribution with shape parameter equal to m times the shape parameter of the single variable. However, this turns out to be a more challenging issue for PH distribution and requires further analysis of the parameters that are in a matrix format.

Clearly, if variable C is the sum of two PH variables A and B, the transition rate matrix of C can be shown as:

$$\mathbb{S}^{(C)} = \begin{pmatrix} \mathbb{S}^{(A)} & \mathbf{S}^0\boldsymbol{\pi}^{(B)} \\ \mathbf{0} & \mathbb{S}^{(B)} \end{pmatrix}, \tag{3.12}$$

and the initial probability vector becomes $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)}, \boldsymbol{\pi}^{(A)}(N+1)\boldsymbol{\pi}^{(B)}]$. The term $\boldsymbol{\pi}^{(A)}(N+1)$ is the probability of process A starting in an absorption state that is considered 0 here, so $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)}, \mathbf{0}_{1\times N}]$. When modeling aggregate failure-time data, the sum of m_k similar PH variables has a transition rate matrix consisting of submatrices equal to the single PH variable transition rate

matrix and failure vector. The design of these matrices is in the following form:

$$\mathbb{S}^{(new)} = \begin{pmatrix} \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & & \ddots & \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbb{S} \end{pmatrix}_{Nm_k \times Nm_k} . \tag{3.13}$$

In the previous EM algorithm, each element estimated in the E-step is comprised of the mean of M matrices driven from the data set. In case of aggregate data, the sizes of matrices contributing to calculating the missing values are different. As such, a different approach should be developed for model parameter estimation.

More specially, matrices $\mathbf{f}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$, $\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$ and $\mathbf{F}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$ have different dimensions for different data points. Indeed, $\mathbf{f}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$ is a $1 \times Nm_k$, $\mathbf{b}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$ is an $Nm_k \times 1$ and $\mathbf{F}_{(\boldsymbol{\pi},\mathbb{S}),t_k}$ is an $Nm_k \times Nm_k$ matrix. Consider $\mathbf{f}$ and $\mathbf{b}$ as m_k concatenated $1 \times N$ matrices, each of which referring to one component failure and representing an N-phase Markov process. Each matrix $\mathbf{F}$ includes m_k diagonal $N \times N$ submatrices, each being equivalent to matrix $\mathbf{F}$ for a single component's failure time. Note that in this method matrix $\mathbf{S}^{0(k)}$ shows the rate that in a corresponding phase a component is moved to the absorption of the m_k'th component. To catch the single component absorption rates, the rates of transition to the phases relative to the other components should be added to $\mathbf{S}^{0(k)}$. The values of $\mathbf{S}^{0(k)}$ that are relative to any component except the m_k'th, are zero. For the proposed EM algorithm, a new absorption rate matrix should be defined, which consists of the individual absorption rates. Each $\mathbf{S}^0\boldsymbol{\pi}$ in Equation (4.9) contributes to absorption rates as

follows:

$$\mathbf{S}^0\boldsymbol{\pi} = \begin{pmatrix} \mathbf{S}^0(1)\boldsymbol{\pi}(1) & \mathbf{S}^0(1)\boldsymbol{\pi}(2) & \cdots & \mathbf{S}^0(1)\boldsymbol{\pi}(N) \\ \mathbf{S}^0(2)\boldsymbol{\pi}(1) & \mathbf{S}^0(2)\boldsymbol{\pi}(2) & \cdots & \mathbf{S}^0(2)\boldsymbol{\pi}(N) \\ \vdots & \vdots & \ddots & \cdots \\ \mathbf{S}^0(N)\boldsymbol{\pi}(1) & \mathbf{S}^0(N)\boldsymbol{\pi}(2) & \cdots & \mathbf{S}^0(N)\boldsymbol{\pi}(N) \end{pmatrix}. \tag{3.14}$$

Consequently, the absorption rate at phase i of each individual component becomes $\mathbf{d}_i = \sum_{j=1}^{N} \mathbf{S}^0(i)\boldsymbol{\pi}(j)$. Thus, the new absorption vector $\mathbf{d}^{(k)}$ could be constructed and used in the E-step as:

$$\mathbf{d}^{(k)} = [\mathbf{d}_1^{(k)}, \cdots, \mathbf{d}_N^{(k)}, \cdots, \mathbf{d}_1^{(k)}, \cdots, \mathbf{d}_N^{(k)}]_{1\times Nm_k} \tag{3.15}$$

Note that $\mathbf{d}^{(k)}$ is not the actual absorption rate matrix of data point k, but it is the set of hidden absorption matrices related to each individual component failed in that data point.

The likelihood function for this case can be described as Equation (4.15) after modifying the definitions for some variables. In particular, $\boldsymbol{\tau} = (t_1, t_2, \cdots, t_M, m_1, m_2, \cdots, m_M)$, and $\mathbb{S}$ represents the single component transition rate matrix. Each single component lifetime is related to one Markov process, and for a data point k, m_k Markov processes occur successively. In addition, the unobserved variables, B_i, Z_i, and N_{ij} for the case of aggregate data are defined as:

B_i: the number of times the Markov process started in phase $lN + i$, $l = 0, \cdots, max(m_k) - 1$.

Z_i: the time that was spent in phase $ln + i$; $\quad l = 0, \cdots, max(m_k) - 1$.

N_{ij}: the number of jumps from phase $lN + i$ to phase $lN + j$; $\quad l = 0, \cdots, max(m_k) - 1$.

Using the above definitions, we can make sure that inside the Markov process of each data point, before reaching the absorption state of the current component, transitions to those phases related

to the subsequent components are not allowed. Then, the equations for the E-step are:

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[B_i] = \frac{1}{M}\sum_{k=1}^{p}\sum_{l=0}^{m_k}\frac{\boldsymbol{\pi}^{(k)}(i+ln)\mathbf{b}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}(i+ln)}{\boldsymbol{\pi}^{(k)}\mathbf{b}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}}, \tag{3.16}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i] = \frac{1}{M}\sum_{k=1}^{p}\sum_{l=0}^{m_k}\frac{\mathbf{F}^{(k)}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}(i+ln,i+ln)}{\boldsymbol{\pi}^{(k)}\mathbf{b}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}}, \tag{3.17}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{ij}] = \frac{1}{M}\sum_{k=1}^{p}\sum_{l=0}^{m_k}\frac{\mathbb{S}^{(k)}(i+ln,j+ln)\mathbf{F}^{(k)}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}(i+ln,j+ln)}{\boldsymbol{\pi}^{(k)}\mathbf{b}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}}, \tag{3.18}$$

$$E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{in+1}] = \frac{1}{M}\sum_{k=1}^{p}\sum_{l=0}^{m_k}\frac{\mathbf{d}^{(k)}(i+ln)\mathbf{f}^{(k)}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}(i+ln)}{\boldsymbol{\pi}^{(k)}\mathbf{b}_{(\boldsymbol{\pi}^{(k)},\mathbb{S}^{(k)}),t_k}}, \tag{3.19}$$

where p is the number of available data points, $M = \sum_{k=1}^{p} m_k$ is the total number of failures, $\boldsymbol{\pi}$ and $\mathbb{S}$ are the estimated initial probability vector and transition rate matrix of a single component's failure time, $\boldsymbol{\pi}^{(k)}$ and $\mathbb{S}^{(k)}$ are those of m_k components and $i, j = 1, 2, \cdots, N$. Using these E-step equations, the M-step can be performed using the formulas stated in Section 4.4.1.1. In summary, the proposed EM algorithm for handling aggregate data is as follows:

(i) Define initial values for the parameters of an N-phase PH distribution.

(ii) Define the proper transition rate and absorption matrices for each data point based on the corresponding number of failures.

(iii) Define the statistics of EM algorithm as in Equation (3.6) separately for each data point.

(iv) Use Equations (4.17) - (4.20) to estimate the unobserved data based on the current parameter estimates.

(v) Use Equation (4.21) to update the parameter estimates.

(vi) If a stopping criterion (e.g., a certain number of iterations or the difference between the likelihood values of the last two iterations) is met, stop. Otherwise, go to step (iv).

3.3.3 Model Selection and Setting of Initial Values

To avoid the non-identifiability problem of parameters, we have used Coxian distribution. It can be shown that any general PH distribution can be represented by a Coxian distribution. Using

Coxian distribution with ordered diagonal values eliminates the redundancy in parameters. As in the EM algorithm, the initial values should be used for the parameters, we suggest an approach to obtain initial parameter values. Based on our experiments, the algorithm is not highly sensitive to the initial parameter values. As long as the initial values are not chosen such that an extremely low likelihood is attained, the algorithm can find its way to the optimum solution. Although this seems like an easy job, in practice, it can be difficult to obtain a reasonable first guess. As Erlang distribution is a special case of Coxian distribution, the parameter estimate of Erlang distribution can be used as the start point. Recall that in the EM algorithm for PH distributions, if a value is initially set to zero, it will be zero in the ML estimate. To avoid all zero values in the absorption rate matrix, a small value relative to the optimum λ of Erlang distribution can be assigned to the absorption rates.

Since the convolution of Erlang distribution is tractable, it can be easily used for aggregate data. If the lifetime of each component follows $Erlang(N, \lambda)$, an aggregate data point with m failures follows $Erlang(mN, \lambda)$. Then, parameter λ can be found by maximizing the likelihood function:

$$\mathcal{L}((\pi, \mathbb{S})|\tau) = \prod_{k=1}^{p} \frac{\lambda^{m_k N} t_k^{m_k N-1} e^{-\lambda t_k}}{(m_k N - 1)!} \tag{3.20}$$

where

$$\mathbb{S} = \begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -\lambda & \lambda \\ 0 & \cdots & \cdots & 0 & -\lambda \end{pmatrix}. \tag{3.21}$$

When fitting a PH distribution to aggregate data, models with different numbers of phases can be considered. Although it can be shown that increasing the number of phases could potentially

improve the likelihood, a model selection method is required to determine the most suitable number of phases in some sense. It is worth pointing out that using Akaike Information Criterion (AIC) may not be effective in selecting a PH distribution, even for Coxian distribution, as the number of phases often grows rapidly in comparison to the likelihood value. In this work, the Maximum a Posteriori (MAP) estimation method with a Laplacian prior is used for the purpose of model selection. Specially, the Laplacian prior is denoted as:

$$p(\boldsymbol{\Theta}|(\boldsymbol{\mu}, \boldsymbol{b})) = \prod_{i=1}^{n} \frac{1}{2b_i} e^{-\frac{|\Theta_i - \mu_i|}{b_i}}, \tag{3.22}$$

where $\boldsymbol{\Theta}$ contains the parameters of the distribution under test, n equals the number of parameters, and $(\boldsymbol{\mu}, \boldsymbol{b})$ is the vector of Laplacian distribution parameters. With the likelihood function $f(\boldsymbol{\tau}|\boldsymbol{\Theta})$ as Equation (4.15) with $\boldsymbol{\tau} = (t_1, t_2, \cdots, t_M, m_1, m_2, \cdots, m_M)$, the MAP estimator is $\text{argmax}_{\boldsymbol{\Theta}} f(\boldsymbol{\tau}|\boldsymbol{\Theta}) p(\boldsymbol{\Theta}|(\boldsymbol{\mu}, \boldsymbol{b}))$, and the candidate distribution with an appropriate number of parameters that results in the highest MAP value will be selected.

3.3.4 ML-based Confidence Interval

In this section, a method for finding the confidence intervals of quantities of interest using Fisher information is presented. It is worth pointing out that Fisher information for PH distribution has been studied in the literature, but it has never been extended and used in dealing with aggregate failure-time data.

Let $\boldsymbol{\Theta} = (\boldsymbol{\pi}, vector(\mathbb{S}))'$ be the vector containing all the parameters to be estimated, $\boldsymbol{\tau}$ be the available aggregate data, and $\ell(\boldsymbol{\Theta}|\boldsymbol{\tau}) = \ln \mathcal{L}(\boldsymbol{\Theta}|\boldsymbol{\tau})$ be the log-likelihood function. The empirical Fisher information matrix can be expressed as:

$$I(\boldsymbol{\Theta}) = -\left[\frac{\partial^2}{\partial \boldsymbol{\Theta} \partial \boldsymbol{\Theta}'} \ell(\boldsymbol{\Theta}|\boldsymbol{\tau})\right]. \tag{3.23}$$

Due to the special structure of PH distribution and the fact that the parameters are masked inside the transition rate matrix, the formation of Fisher information matrix is not straightforward. Bladt et al.

(2011) proposed an EM algorithm and a Newton-Raphson method to attain the Fisher information matrix for the parameters of PH distribution. In this work, we will use the Newton-Raphson method for Fisher information matrix estimation and extend their method to deal with aggregate data. To this end, some of the expressions used in their method need to be updated.

First, the Newton-Raphson method is explained here. The derivative of the log-likelihood function with respect to the vector of parameters is:

$$\frac{\partial \ell(\mathbf{\Theta}|\boldsymbol{\tau})}{\partial \mathbf{\Theta}} = \sum_{k=1}^{M} \frac{1}{f(t_k|\mathbf{\Theta})} \frac{\partial f(t_k|\mathbf{\Theta})}{\partial \mathbf{\Theta}}, \tag{3.24}$$

where $f(\cdot)$ is the PDF. Note that taking the derivative of PDF f with respect to $\mathbf{\Theta}$ is an issue, for which the following formulas are created. The parameters of an N-phase PH distributions are $N-1$ elements of $\boldsymbol{\pi}$, non-diagonal elements of $\mathbb{S}$, noted as d_{hn}, and all the elements of $\mathbf{S}^0$, noted as d_h, for $h, n = 1, 2, \cdots, N$. To get started, using uniformization for matrix exponential, we define $c = max\{-d_{hh} : 1 < h < N\}$ and $\mathbf{K} = (1/c)\mathbb{S} + \mathbf{I}$, where $\mathbf{I}$ is the identity matrix. Then, we have:

$$e^{\mathbb{S}t} = \sum_{r=0}^{\infty} e^{-ct} \frac{ct^r}{r!} \mathbf{K}^r. \tag{3.25}$$

Let $\mathbf{\Psi}(t) = e^{(\mathbb{S}t)}$. By decomposing $\boldsymbol{\pi}$ into $\sum_{j=1}^{p-1} \pi_j \mathbf{e}_j^\top + (1 - \sum_{j=1}^{p-1} \pi_j)\mathbf{e}_N^\top$, we have:

$$\frac{\partial f(t|\mathbf{\Theta})}{\partial \pi_h} = \mathbf{e}_h^\top \mathbf{\Psi}(t)\mathbf{S}^0 - \mathbf{e}_N^\top \mathbf{\Psi}(t)\mathbf{S}^0, \tag{3.26}$$

$$\frac{\partial f(t|\mathbf{\Theta})}{\partial d_{hn}} = \boldsymbol{\pi} \frac{\partial \mathbf{\Psi}(t)}{\partial d_{hn}} \mathbf{S}^0, \qquad h \neq n, \tag{3.27}$$

$$\frac{\partial f(t|\mathbf{\Theta})}{\partial t_h} = \boldsymbol{\pi} \mathbf{\Psi}(t)\mathbf{e}_h + \boldsymbol{\pi} \frac{\partial \mathbf{\Psi}(t)}{\partial d_h} \mathbf{S}^0, \tag{3.28}$$

where $\mathbf{e}_j$ is a column vector with 1 in the jth place and 0 elsewhere. Then, the problem is reduced

to calculating the partial derivatives of $\mathbf{\Psi}(t)$:

$$\frac{\partial \mathbf{\Psi}(t)}{\partial \Theta_q} = e^{-ct} \sum_{s=0}^{\infty} \frac{(ct)^{s+1}}{(s+1)!} \mathbf{D}_q(s) + \frac{\partial c}{\partial \Theta_q} t e^{\mathbb{S}t} (\mathbf{K} - \mathbf{I}), \quad q = 1, 2, \cdots, N^2 + N - 1, \tag{3.29}$$

where $\mathbf{D}_q(s) = \partial \mathbf{K}^{s+1} / \partial \Theta_q$. To calculate this, partial derivatives of powers of the transition rate matrix with respect to the parameters are required. Especially, we have:

$$\frac{\partial \mathbb{S}^r}{\partial \Theta_q} = \sum_{k=0}^{r-1} \mathbb{S}^k \frac{\partial \mathbb{S}}{\partial \Theta_q} \mathbb{S}^{r-1-k}, \tag{3.30}$$

where $[\partial \mathbb{S} / \partial d_{ij}]_{ij} = 1$, $[\partial \mathbb{S} / \partial d_{ij}]_{ii} = -1$, $[\partial \mathbb{S} / \partial d_i]_{ii} = -1$, and the rest of the elements are all 0. Based on these basic definitions, the following results can be obtained:

$$\frac{\partial^2 \mathbb{S}^r}{\partial \Theta_p \partial \Theta_q} = \sum_{k=0}^{r-1} \mathbb{S}^k \frac{\partial \mathbb{S}}{\partial \Theta_q} \frac{\partial \mathbb{S}^{r-1-k}}{\partial \Theta_p} + \mathbb{S}^{r-1-k} \frac{\partial \mathbb{S}^k}{\partial \Theta_p} \frac{\partial \mathbb{S}}{\partial \Theta_q}, \tag{3.31}$$

$$\frac{\partial^2 e^{\mathbb{S}t}}{\partial \Theta_p \partial \Theta_q} = e^{-ct} \sum_{k=0}^{\infty} \frac{(ct)^{k+1}}{(k+1)!} \frac{\partial^2 \mathbf{K}^{k+1}}{\partial \Theta_p \partial \Theta_q} + \frac{\partial c}{\partial \Theta_q} t \Big(e^{\mathbf{T}t} \frac{\partial \mathbf{K}}{\partial \Theta_p} + \frac{\partial e^{\mathbf{T}t}}{\partial \Theta_p} (\mathbf{K} - \mathbf{I}) \Big), \tag{3.32}$$

where $p, q = 1, 2, \cdots, N^2 + N - 1$. It is worth pointing out that these formulas are used to produce a Fisher information matrix based on individual failure-time data (Bladt et al., 2011). In this work, these formulas are extended to adapt to aggregate failure-time data.

The second derivative of log-likelihood function is:

$$\frac{\partial^2 \ell(\mathbf{\Theta} | \boldsymbol{\tau})}{\partial \mathbf{\Theta} \partial \mathbf{\Theta'}} = \sum_{k=1}^{m} \frac{1}{f(t_k | \mathbf{\Theta})^2} \left[f(t_k | \mathbf{\Theta}) \frac{\partial^2 f(t_k | \mathbf{\Theta})}{\partial \mathbf{\Theta} \partial \mathbf{\Theta'}} - \frac{\partial f(t_k | \mathbf{\Theta})}{\partial \mathbf{\Theta}} \frac{\partial f(t_k | \mathbf{\Theta})}{\partial \mathbf{\Theta'}} \right]. \tag{3.33}$$

For each aggregate failure-time data t_k, the corresponding PDF is defined based on the number of failed components in that data point, as shown previously. In other words, we have a different PDF (thus a different transition rate matrix and different number of phases), as given in Equation (4.9), for each data point as:

$$f_k(t | (\boldsymbol{\pi}, \mathbb{S})) = \boldsymbol{\pi}_k e^{\mathbb{S}_k} \mathbf{S}_k^0. \tag{3.34}$$

Moreover, $\partial\mathbb{S}/\partial\Theta_q$ should be updated. For data point k with m_k failures, $[\partial\mathbb{S}/\partial d_{hn}]_{uv} = 1$ and $[\partial\mathbb{S}/\partial d_{hn}]_{uu} = -1$, where $h, n = 1, 2, \cdots, N$, $h \neq n$, $u = h + rN$, $v = n + rN$, $r = 0, 1, \cdots, N-1$. The rest of the parameters of this $Nm_k \times Nm_k$ matrix will be 0s. The following example shows the derivative of transition rate matrix with respect to a parameter when $N = 3$ and two cumulative failures:

$$\frac{\partial\mathbb{S}_k}{\partial d_{23}} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}. \tag{3.35}$$

Note that in this case, the only independent parameters are again the parameters of the PH distribution for a single component, and the total number of parameters is $N^2 + N - 1$. So, for each data point, no matter how many failures it contains, $\boldsymbol{\Theta}$ always contains $N^2 + N - 1$ model parameters.

After producing the Fisher information matrix, the normal approximation method can be used to attain the confidence intervals of interest. The Wald statistic is defined as:

$$W = (\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})'[\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\Theta}}}]^{-1}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}), \tag{3.36}$$

where $\hat{\boldsymbol{\Theta}}$ is the MLE of $\boldsymbol{\Theta}$, and $\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\Theta}}}$ is the estimated variance-covariance matrix obtained by taking the inverse of Fisher information matrix. Since $\hat{\boldsymbol{\Theta}}$ asymptotically follows a multivariate Normal distribution with parameters $\boldsymbol{\Theta}$ and $\boldsymbol{\Sigma}$, W follows a Chi-square distribution with degrees of freedom equal to the length of $\boldsymbol{\Theta}$ (i.e., the number of parameters noted as v). Then, a $100(1-\alpha)\%$ approximate confidence region for $\boldsymbol{\Theta}$ can be obtained from:

$$(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta})'[\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\Theta}}}]^{-1}(\hat{\boldsymbol{\Theta}} - \boldsymbol{\Theta}) \leq \chi^2_{(1-\alpha;v)}. \tag{3.37}$$

In certain circumstances, the statistics of Wald and likelihood ratio are equivalent, so that the distribution is exact. For other cases, it can be shown that Wald interval is the quadratic approximation to a likelihood-based confidence region (Meeker and Escobar (1995)). Regarding PH distributions, this is an asymptotic approximate method, and exact pivotal quantities for the parameters are not discussed in the literature. In this chapter, we provide the ML confidence interval for each individual parameter by Normal approximation. For example, for transition rate parameter λ_1, we have $[\underset{\sim}{\lambda_1}, \tilde{\lambda_1}] = \left[\hat{\lambda_1}/V, \hat{\lambda_1} \times V\right]$, where $V = \exp\left(z_{1-\alpha/2}\hat{se}_{\hat{\lambda_1}}/\hat{\lambda_1}\right)$, $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of the standard Normal distribution, and $\hat{se}_{\hat{\lambda_1}} = \sqrt{\widehat{Var}(\hat{\lambda_1})}$.

For practical purposes, the confidence interval for the CDF of failure-time distribution is of interest. Based on the estimated variance-covariance matrix, the confidence interval for the CDF can be obtained as follows. First, the variance of the CDF estimate is calculated via the delta method as:

$$\widehat{Var}(F(t)) = \left[\frac{\partial F}{\partial \mu_1}, \cdots, \frac{\partial F}{\partial \lambda_2}\right] \hat{\mathbf{\Sigma}}_{\hat{\mathbf{\Theta}}} \left[\frac{\partial F}{\partial \mu_1}, \cdots, \frac{\partial F}{\partial \lambda_2}\right]'. \tag{3.38}$$

Then, the approximate confidence interval for the CDF can be expressed as:

$$[\underset{\sim}{F}, \tilde{F}] = \left[\frac{\hat{F}(t)}{\hat{F}(t) + (1-\hat{F}(t)) \times W}, \frac{\hat{F}(t)}{\hat{F}(t) + (1-\hat{F}(t))/W}\right], \tag{3.39}$$

where

$$W = \exp\left(\frac{z_{1-\alpha/2}\hat{se}_{\hat{F}}}{\hat{F}(t_e)(1-\hat{F}(t_e))}\right) \text{ and } \hat{se}_{\hat{F}} = \sqrt{\widehat{Var}(\hat{F}(t))}. \tag{3.40}$$

3.4 Bayesian Alternative

A Bayesian alternative is also provided in this work for reliability estimation using aggregate data. The studies by Ausín et al. (2008) and McGrory et al. (2009) concentrated on Bayesian methods for Coxian distributions. In this section, we will use the method developed by McGrory et al.

(2009) and extend their model to estimate the parameters of Coxian based on aggregate data. Moreover, using the same posterior distribution for Coxian and with the assistance of Metropolis-Hastings algorithm, credible intervals are estimated for the model parameters.

McGrory et al. (2009) introduced a Bayesian formulation for a Coxian distribution with covariates and unknown number of phases. They considered a Gamma prior distribution for each parameter. Here, we will utilize the same model while ignoring covariates. For the transition rate matrix of an N-phase Coxian distribution given in Equation (4.3), we assume that the prior distributions of model parameters are $\lambda_j \sim Gamma(\alpha_j, \beta_j)$, $j = 1, 2, \cdots, N-1$, and $\mu_j \sim Gamma(\gamma_j, \sigma_j)$, $j = 1, 2, \cdots, N$ (McGrory et al. (2009)). Then, the posterior distribution can be obtained as:

$$\begin{aligned} p(\mathbf{\Theta}_N, N|\mathbf{y}) &\propto p(\mathbf{y}|\mathbf{\Theta}_N, N)p(\mathbf{\Theta}_N|N)p(N) \\ &= \prod_{i=1}^{M} \pi e^{(\mathbb{S}_i y_i)} \mathbf{S}_i^0 \prod_{j=1}^{N-1} \frac{1}{\Gamma(\alpha_i)} \frac{1}{\beta_j^{\alpha_j}} \lambda_j^{\alpha_j - 1} exp(-\frac{\lambda_j}{\beta_j}) \\ &\quad \times \prod_{j=1}^{N} \frac{1}{\Gamma(\gamma_j)} \frac{1}{\sigma_j^{\gamma_j}} \mu_j^{\gamma_j - 1} exp(-\frac{\mu_j}{\sigma_j}) \times p(N). \end{aligned} \tag{3.41}$$

For the case of aggregate data, λ_j and μ_j are the parameters of Coxian distribution for a single component failure, but $\mathbb{S}_i$ and $\mathbf{S}_i^0$ are those related to data point i based on Equation (4.9), which have different dimensions for different data points. Note that a subscript N is added to the parameter vector $\mathbf{\Theta}$ to emphasize the number of phases of the current model, which may be adjusted.

RJMCMC (Green (1995)) is a method that enables jumps between models with different dimensions. The algorithm proposed by McGrory et al. (2009) considers three main possibilities with equal probabilities: a fixed dimension update of the parameters, splitting the phase into two or combining two existing phases into one, and birth of a new phase or death of an existing phase. In particular, fixed dimension parameter update is done through a Metropolis-Hastings algorithm. For dimension changing reversible jump moves, some basic definitions are needed. Let the current number of phases be N and the proposed number of phases be N^*. In each jump step, the dimension can only increase or decrease by one unit while satisfying the requirement on the maximum

and minimum numbers of phases. u, v, u^* and v^* are auxiliary variables defined to keep the dimensionality of the current and proposed parameter spaces, (Θ_N, u, v) and (Θ_{N^*}, u^*, v^*), respectively. We define:

$$R = \frac{p(\mathbf{y}|\mathbf{\Theta}_{N^*}, N^*)p(\mathbf{\Theta}_{N^*})p(N^*)}{p(\mathbf{y}|\mathbf{\Theta}_N, N)p(\mathbf{\Theta}_N)p(N)} \times \frac{Q_{N^*,N}p(u^*, v^*|N^*, N, \mathbf{\Theta}_{N^*})}{Q_{N,N^*}p(u, v|N, N^*, \mathbf{\Theta}_N)} \times \left|\frac{\partial(\mathbf{\Theta}_{N^*}, u^*, v^*)}{\partial(\mathbf{\Theta}_N, u, v)}\right|, \quad (3.42)$$

where Q_{N,N^*} is the probability of moving from N to N^*, and the third term is the Jacobian for transformation, which will be addressed later. Then, the probability of accepting a proposed move is $min(R, 1)$. To perform a reasonable mapping, it is ensured that the mean time and probability of absorption in current and proposed phase(s) remain similar, such that:

$$\frac{\mu}{\mu + \lambda} = \frac{\mu_a}{\mu_a + \lambda_a} + \big(\frac{\lambda_a}{\mu_a + \lambda_a} \times \frac{\mu_b}{\mu_b + \lambda_b}\big), \quad (3.43)$$

$$\frac{\mu}{\mu + \lambda} = \frac{1}{\mu_a + \lambda_a} + \frac{1}{\mu_b + \lambda_b}. \quad (3.44)$$

For split and birth moves, where one new phase is introduced, μ and λ denote the rates before transformation, and μ_a, μ_b, λ_a and λ_b denote the rates after the transformation. For combine and death moves, the process will be performed reversely.

Accordingly, for each move we need to find the new parameters, based on Equations (4.28) and (3.44), as well as the Jacobian of the transformation, which will be succinctly described here. Note that split and combine moves cannot be applied for the final phase, while birth and death moves are only performed on the final phase.

Via combine move: $(\mu_a, \lambda_a, \mu_b, \lambda_b) \rightarrow (u, v, \mu, \lambda)$, where $u = \mu_a$ and $v = \lambda_a$, we have:

$$\mu = \frac{\mu_a\mu_b + \mu_a\lambda_b + \lambda_a\mu_b}{\mu_a + \lambda_a + \mu_b + \lambda_b}, \quad (3.45)$$

$$\lambda = \frac{\lambda_a\lambda_b}{\mu_a + \lambda_a + \mu_b + \lambda_b}, \quad (3.46)$$

$$|J| = \frac{(\mu_a + \lambda_a)^2}{(\mu_a + \lambda_a + \mu_b + \lambda_b)^3}. \quad (3.47)$$

Via split move: $(u, v, \mu, \lambda) \rightarrow (\mu_a, \lambda_a, \mu_b, \lambda_b)$, again $u = \mu_a$ and $v = \lambda_a$, and u and v should be

simulated from $u \sim N_T(2\mu, \sigma^2)$ and $v \sim N_T(2\lambda, \sigma^2)$ truncated at 0. The Jacobian for a split move is the reciprocal of the one for a combine move:

$$\mu_b = \frac{\mu_a^2\lambda + \mu_a\lambda_a\lambda - \lambda_a\mu\mu_a - \lambda_a^2\mu}{\lambda_a(-\mu_a - \lambda_a + \mu + \lambda)}, \tag{3.48}$$

$$\lambda_b = -\frac{(\mu_a + \lambda_a)^2\lambda}{\lambda_a(-\mu_a - \lambda_a + \mu + \lambda)}. \tag{3.49}$$

Via death move: $(\mu_a, \lambda_a, \mu_b) \to (u, v, \mu)$, we have:

$$\mu = \frac{(\mu_a + \lambda_a)\mu_b}{(\mu_b + \mu_a + \lambda_a)}, \tag{3.50}$$

$$|J| = \frac{(\mu_a + \lambda_a)^2}{(\mu_b + \mu_a + \lambda_a)^2}. \tag{3.51}$$

Via birth move: $(u, v, \mu) \to (\mu_a, \lambda_a, \mu_b)$ with u and v being simulated from $u \sim N_T(\mu, \sigma^2)$ and $v \sim N_T(\mu, \sigma^2)$ truncated at 0, the Jacobian is the reciprocal of the expression used for death move and

$$\mu_b = \frac{(u + v)\mu}{u + v - \mu}. \tag{3.52}$$

For more detailed explanations of the RJMCMC method, readers are referred to McGrory et al. (2009). This method can be used for updating the parameters of Coxian distribution for a single component as a part of sum of a number of variables with the use of posterior distribution stated in Equation (4.25). As the RJMCMC method jumps between models with different numbers of phases, model selection is automatically performed within the estimation procedure.

For credible interval estimation, we propose to apply the same Gamma prior distributions for the parameters. The posterior distribution takes the number of phases as constant. Based on the posterior distribution in Equation (4.25) and using Metropolis-Hastings approach to generate parameter estimates, credible intervals for parameters using quantiles of the generated parameter values can be obtained. Needless to say, this model, if used only for credible interval estimation and not for RJMCMC, can be easily extended to handle general PH distributions.

3.5 Numerical Examples

3.5.1 A Simulation Study

3.5.1.1 Capability of Fitting Different Failure-time Distributions

To demonstrate the capability and flexibility of our proposed methods in reliability estimation, aggregate data from different probability distributions are generated, and Coxian distributions are used to fit the data and compared against the true distributions from which the data are generated. In particular, the ML estimation method is illustrated in this study.

The simulated data are generated from Gamma, IG and Weibull distributions. For each distribution, two cases are considered. The first case considers 6 aggregate data points with vector of number of failures $\mathbf{M} = [2 \quad 9 \quad 8 \quad 8 \quad 6 \quad 5]$, and the second case involves 12 aggregate data points with $\mathbf{M} = [2 \quad 2 \quad 9 \quad 9 \quad 8 \quad 8 \quad 8 \quad 8 \quad 6 \quad 6 \quad 5 \quad 5]$. When fitting the Coxian distributions, only the aggregate data are used. However, if individual failure-times are available, to estimate the model parameters using the ML method, Equations (3.7(-(3.11) should be applied. To visualize the estimation capability of the proposed method, the CDF's of the true distribution and the estimated Coxian distribution are shown together in each figure. Moreover, we have saved individual failure times so that the Kaplan-Meier estimate is also calculated and presented in the same figure for comparison.

The distributions are chosen in different ranges for fair comparison. Figure 3.4 shows the results for the aggregate data generated from $Gamma(2.5, 4)$. The result in the left figure is obtained based on 6 data points involving a total of 38 component failures, and the right figure is based on 12 data points for a total of 76 component failures. One can see that the Coxian distribution can mimic the true distribution quite closely, and as the number of data points increases, the deviation from the true distribution becomes negligible. This result illustrates the flexibility of Phase-type distribution in approximating other distributions. For the IG distribution as illustrated in Figure 3.5, the data generated from $IG(10, 8)$ is used. Our results show that the Coxian distribution can also mimic the IG distribution closely. For the two-parameter-Weibull distribution, $Weibull(1, 1.5)$, although

the estimated 3-phase Coxian distribution is a little off from the true distribution, the number of phases can be increased to increase the accuracy. As an illustration, a 6-phase Coxian distribution is used to fit the same Weibull data. Figure 3.7 shows clear improvement by increasing the number of phases. It is worth pointing out that for all the tested distributions, the same number of hidden failures and the same number of simulated data points are used in each case. The flexibility of PH distribution and its capability to handle aggregate data are obvious. The proposed method has potential to be applied for the analysis of aggregate or individual failure-time data when the underlying distribution cannot be conjectured. Moreover, increasing either the number of data points or the number of phases will improve the estimation accuracy of the proposed method. This is particularly favorable for aggregate data, since many probability distributions are intractable for aggregate data.

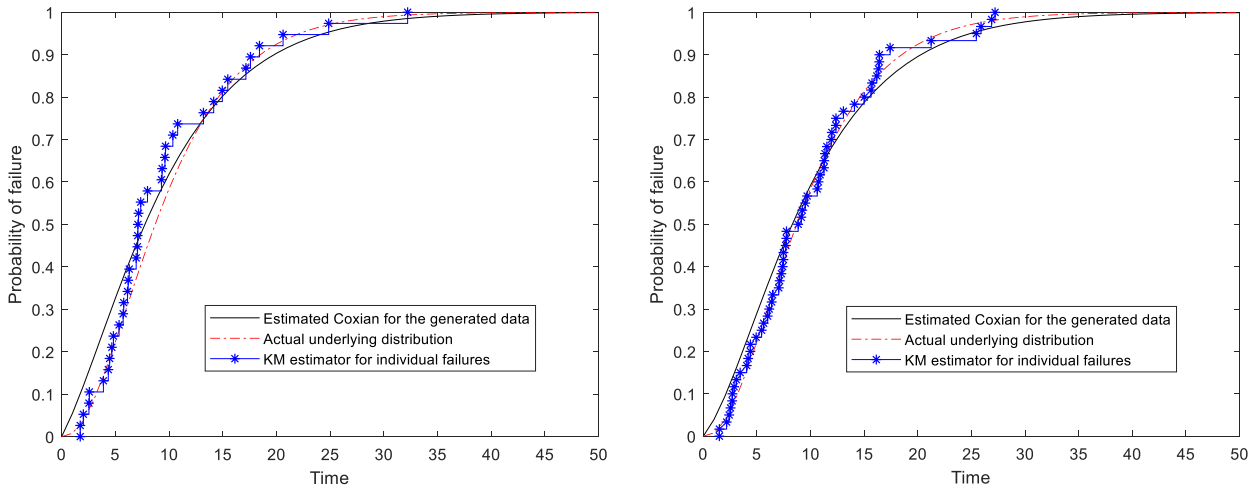

Figure 3.4: Estimated 3-phase Coxian distribution vs. the real underlying distribution, $Gamma(2.5, 4)$, and Kaplan-Meier estimate. The left figure is the result based on 6 data points, and the right figure is based on 12 aggregate data points.

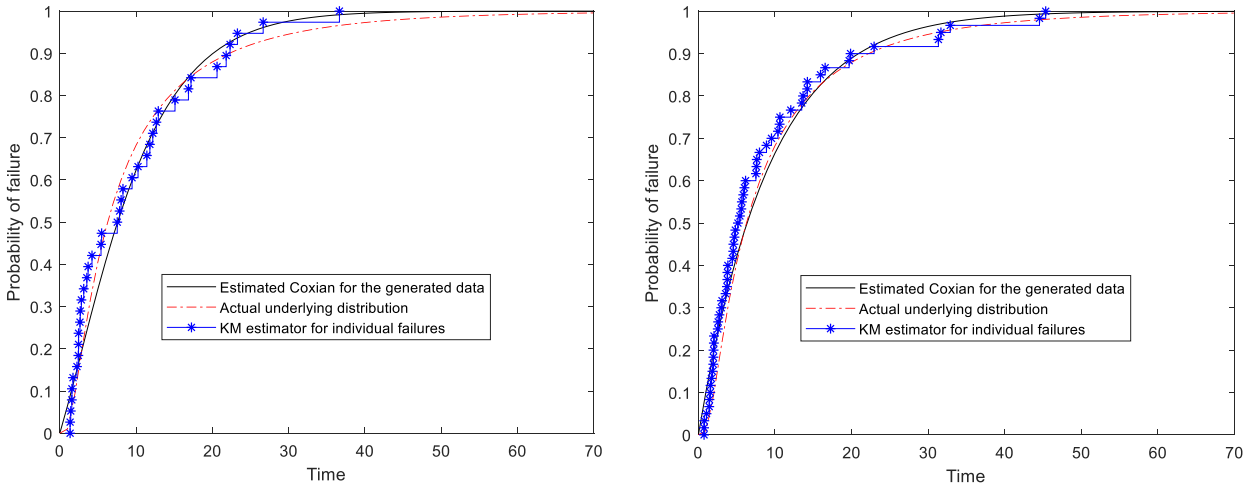

Figure 3.5: Estimated 3-phase Coxian distribution vs. the real underlying distribution, $IG(10, 8)$, and Kaplan-Meier estimate. The left figure is the result based on 6 data points and the right figure is based on 12 aggregate data points.

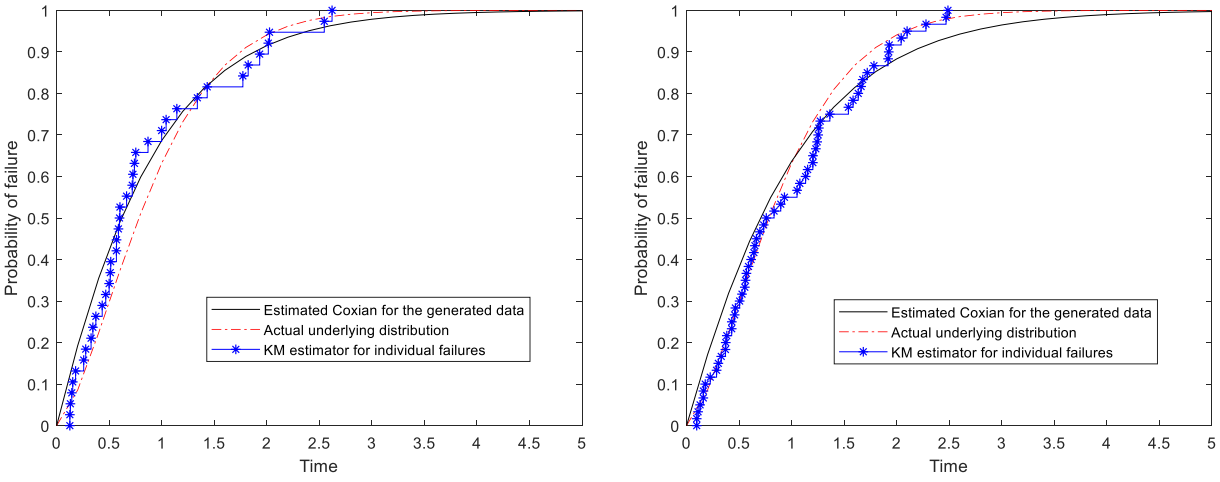

Figure 3.6: Estimated 3-phase Coxian distribution vs. the real underlying distribution, $Weibull(1, 1.5)$, and Kaplan-Meier estimate. The left figure is the result based on 6 data points and the right figure is based on 12 aggregate data points.

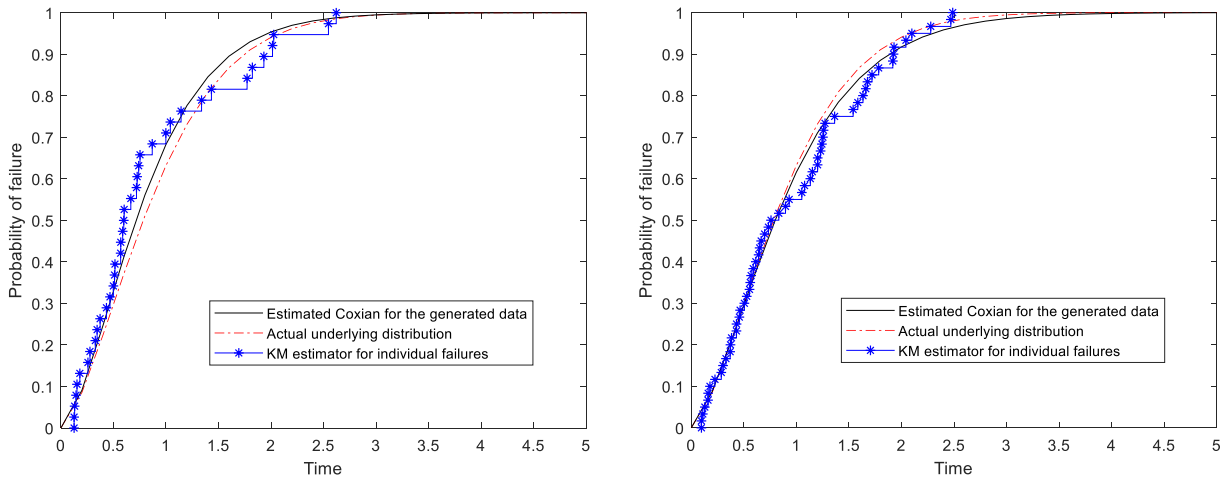

Figure 3.7: Estimated 6-phase Coxian distribution vs. the real underlying distribution, $Weibull(1, 1.5)$, and Kaplan-Meier estimate. The left figure is the result based on 6 data points and the right figure is based on 12 aggregate data points.

3.5.1.2 Study on the Coverage Probability of Normal Approximate Conference Interval

Nonparametric bootstrapping, while being widely used in many situations for interval estimation, should be applied carefully. In particular, the coverage probabilities can be significantly lower than the intended confidence level for small to moderate samples. The reason mainly lies behind the resampling of the bootstrap procedure (Schenker, 1985). In case of aggregate data, failure-times are aggregated into one data point, so practically, we are sampling groups of failures, where the groups do not change. As a result, the resampling problem deteriorates for aggregate data, making the coverage probability of the confidence interval even lower. In this section, the coverage probability of confidence interval obtained using the proposed Normal approximation method is studied against the nonparametric bootstrap method. Note that the coverage probability of credible interval obtained using the Bayesian alternative depends on the selection of prior distribution, thus is not studied in this chapter.

For illustration, a 3-phase Coxian distribution is used to estimate the CDF of a true failure-time distribution, $Weibull(15, 0.95)$. The study is conducted for cases with 6 and 12 aggregate data points, respectively. For each case, the coverage probabilities of the two methods are estimated based on 5000 simulation runs. Table 3.1 shows the results, which clearly show that the bootstrap

CIs for both cases give a much lower coverage probabilities than expected. On the other hand, the proposed Normal approximation method provides much better coverage for those failure-time percentiles.

Table 3.1: Coverage probabilities of 90% CIs using normal approximation and non-parametric bootstrap

Percentile	Normal approx. 6 data points	Normal approx. 12 data points	Bootstrap 6 data points	Bootstrap 12 data points
10	0.8246	0.8751	0.1584	0.6634
50	0.9980	0.9990	0.7426	0.6733
90	0.9965	0.9985	0.6040	0.6832

3.5.2 A Real-world Application

3.5.2.1 The Data

The Reliability Information Analysis Center (RIAC) is a U.S. DoD center who serves for collecting reliability data of fielded systems. Due to the possibilities and technical obstacles in practice, a large amount of the data are not individual component failure-time data (Coit and Jin (2000)). The reliability data shown in Table 3.2 is gathered by RIAC from aircraft indicator lights and has been previously studied by Coit and Jin (2000), and Chen and Ye (2017). In this data, 6 systems were observed, and the number of failures and the cumulative operating time up to the last failure for each system was recorded.

Table 3.2: Aircraft indicator lights failure data

System number	Cumulative operating time (hours)	Number of component failures
1	51000	2
2	194900	9
3	45300	8
4	112400	8
5	104000	6
6	44800	5

For each system k, the cumulative operating time t_k represents the time from the installation of the first component to the failure of the m_k-th component at a certain component position in

system k. For this set of data, the reliability of an individual aircraft indicator light is desired.

3.5.2.2 Reliability Estimation and Model Selection

In this section, the proposed methods are applied on the data set and compared with the three distributions previously studied: Gamma, IG and Normal (Chen and Ye, 2017). The algorithms were run on a computer with Core(TM) i5-6300HQ CPU, 8.00 GB RAM and on Matlab 2017b.

First, the proposed ML estimation method with the new EM algorithm is implemented. Figure 3.8 illustrates the estimated 3-phase Coxian distribution in comparison to the estimated Gamma, IG and Normal distributions studied by Chen and Ye (2017). While the Normal distribution does not provide a very good estimate because of high coefficient of variation. The CDF estimate from the Coxian distribution is close to those of Gamma and IG. The computational time of this method is 57.19 seconds with resulting likelihood value of -30.9821.

For the aircraft indicator light data, using $Laplace(0, 1)$ as the prior distribution, Figure 3.9 shows the MAP estimation result for Coxian distributions with 1 to 10 phases. It is clear that a 3-phase Coxian distribution is suggested. Therefore, the CDF estimate based on the 3-phase Coxian distribution presented in Figure 3.8 is an adequate estimate.

Regarding the Bayesian alternative, the time elapsed for 100 iterations of RJMCMC algorithm with a 20-iteration Metropolis-Hastings, varies between 45 to 50 seconds depending on the size of matrices that are randomly chosen in the algorithm for calculations. Since RJMCMC algorithm moves forward based on random movements to improve the estimation and the number of times each movement is performed during one implementation is different, it could result in different numbers of phases and transition rate matrices of different sizes. The acceptance rate of RJMCMC is around 26% after convergence. The following two matrices are the estimated transition rate matrices from two different implementations of the algorithm:

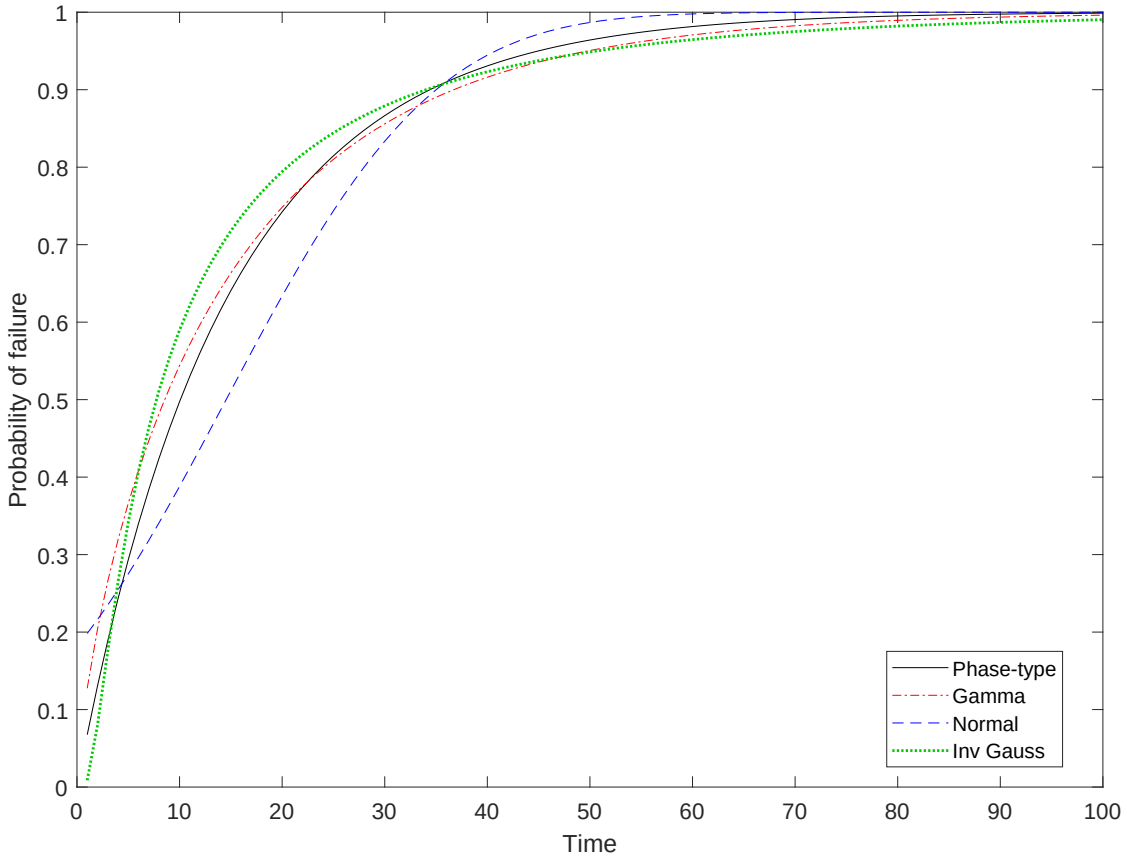

Figure 3.8: CDF's of Gamma, IG, Normal and 3-phase Coxian distributions estimated from the aggregate aircraft indicator light data

$$\mathbb{S}_1 = \begin{pmatrix} -.0674 & 0.0000 & 0 & 0 & 0 & 0 \\ 0 & -0.0233 & 0.0000 & 0 & 0 & 0 \\ 0 & 0 & -0.0072 & 0.0069 & 0 & 0 \\ 0 & 0 & 0 & -0.0001 & 0.0000 & 0 \\ 0 & 0 & 0 & 0 & -0.0516 & 0.0041 \\ 0 & 0 & 0 & 0 & 0 & -0.3830 \end{pmatrix} \text{(6-phase Coxian)},$$

$$\mathbb{S}_2 = \begin{pmatrix} -0.0664 & 0.0000 & 0 \\ 0 & -0.1525 & 0.1099 \\ 0 & 0 & -0.4036 \end{pmatrix} \text{(3-phase Coxian)}.$$

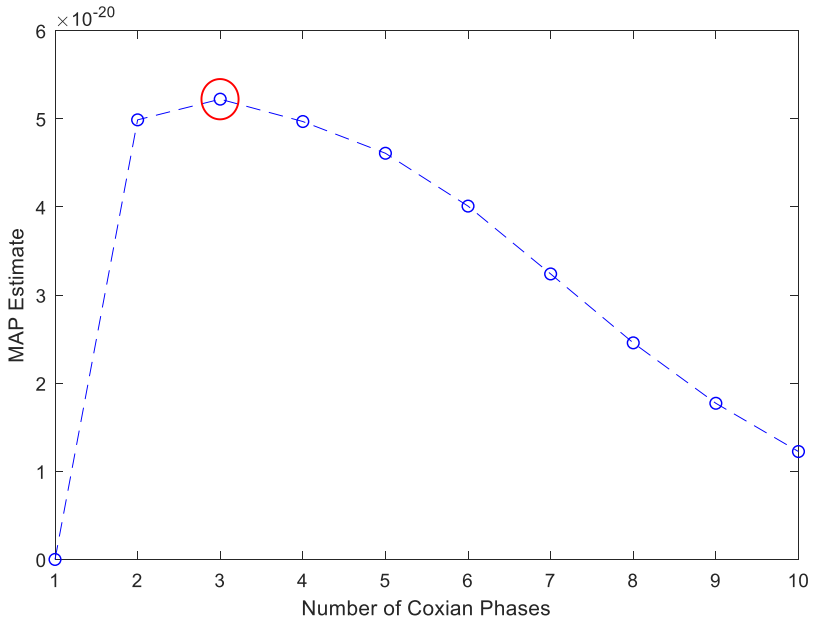

Figure 3.9: MAP model selection method performed for the data provided in Table 3.2 over the range of 1-phase through 10-phase Coxian with Laplacian prior distributions with parameters $(0, 1)$. The maximum MAP estimation suggests a 3-phase Coxian.

Clearly, the two matrices are associated with two different Coxian distributions with different numbers of phases. Unlike the MAP method used in MLE, the disadvantage of this automatic model selection method is that it may not result in a unique model. However, as shown in Figure 3.10, the resulting CDF's obtained from the two implementations are quite close.

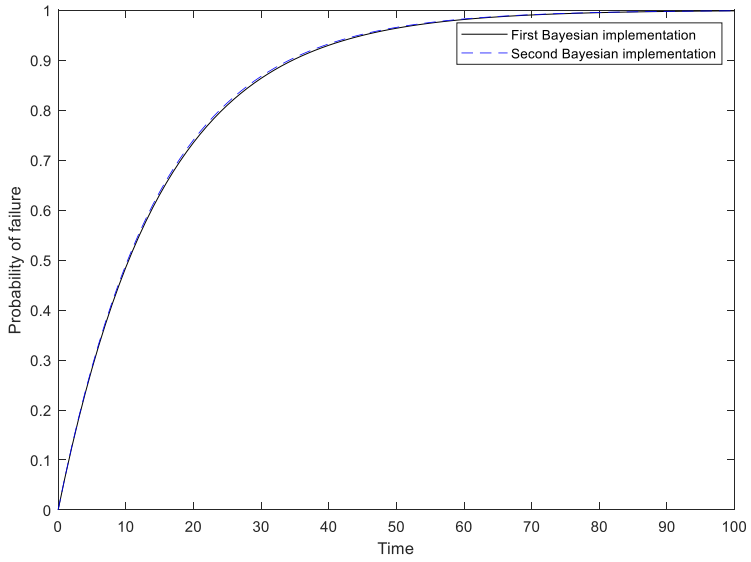

Figure 3.10: CDF estimates of aircraft indicator light from two implementations of the proposed Bayesian method

3.5.2.3 Interval Estimation In this section, the ML confidence intervals (Normal approximation) and Bayesian credible intervals of model parameters and CDF of the 3-phase Coxian distribution are calculated. In particular, the ML confidence interval is found by deriving the Fisher information matrix first followed by calculating the estimated variance covariance matrix as:

$$\boldsymbol{\Sigma}_{\hat{\boldsymbol{\Theta}}} = \begin{pmatrix} 0.0046 & 0.0069 & -0.0052 & -0.0018 & -0.0026 \\ 0.0069 & 0.0113 & -0.0064 & -0.0035 & -0.0043 \\ -0.0052 & -0.0046 & 0.0157 & -0.0084 & -0.0053 \\ -0.0018 & 0.0035 & -0.0084 & 0.0237 & 0.0072 \\ -0.0026 & -0.0043 & -0.0053 & 0.0072 & 0.0206 \end{pmatrix}.$$

Afterwards, the Normal-approximation confidence intervals of model parameters are calculated as addressed in Section 3. Regarding the Bayesian credible intervals of the parameters, the results are obtained based on 1000 Metropolis-Hastings samples. The resulting 90% confidence intervals and credible intervals of model parameters are shown in Table 3.3. The results of the two methods are relatively close except the upper-bound for λ_1.

Table 3.3: C.I.'s based on the MLE and Bayesian methods

Parameter	ML Estimate	90% MLE C.I.	90% Bayesian C.I.
μ_1	0.0702	(0.0274,0.1130)	(0.0514, 0.0902)
μ_2	0.0431	(0, 0.0300)	(0.0014, 0.0529)
μ_3	0.0823	(0, 0.3768)	(0, 0.3861)
λ_1	0.0121	(0, 0.5992)	(0, 0.0336)
λ_2	0.0392	(0, 0.4316)	(0, 0.3525)

Figure 3.11 shows the 90% confidence interval of CDF estimated using Equation (3.39). A nonparametric 90% bootstrap confidence interval is also calculated and provided in Figure 3.12. One can see that the nonparametric bootstrap confidence interval appears to be much narrower than the one from the Normal-approximation alternative. Finally, Figure 3.13 presents the credible interval of CDF from the Bayesian alternative, which depends on the selection of prior distribution and the sample size.

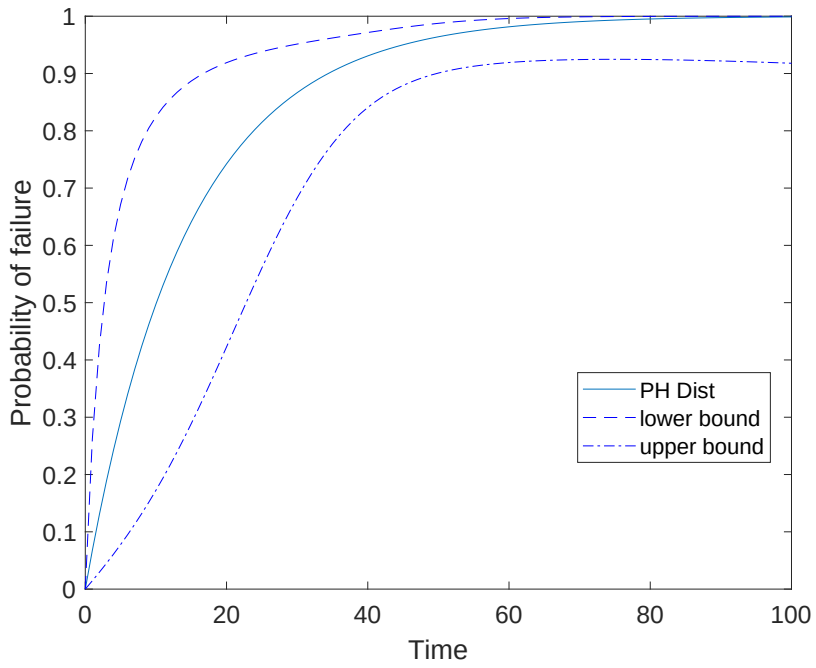

Figure 3.11: 90% Normal approximate confidence interval of CDF based on the 3-phase Coxian

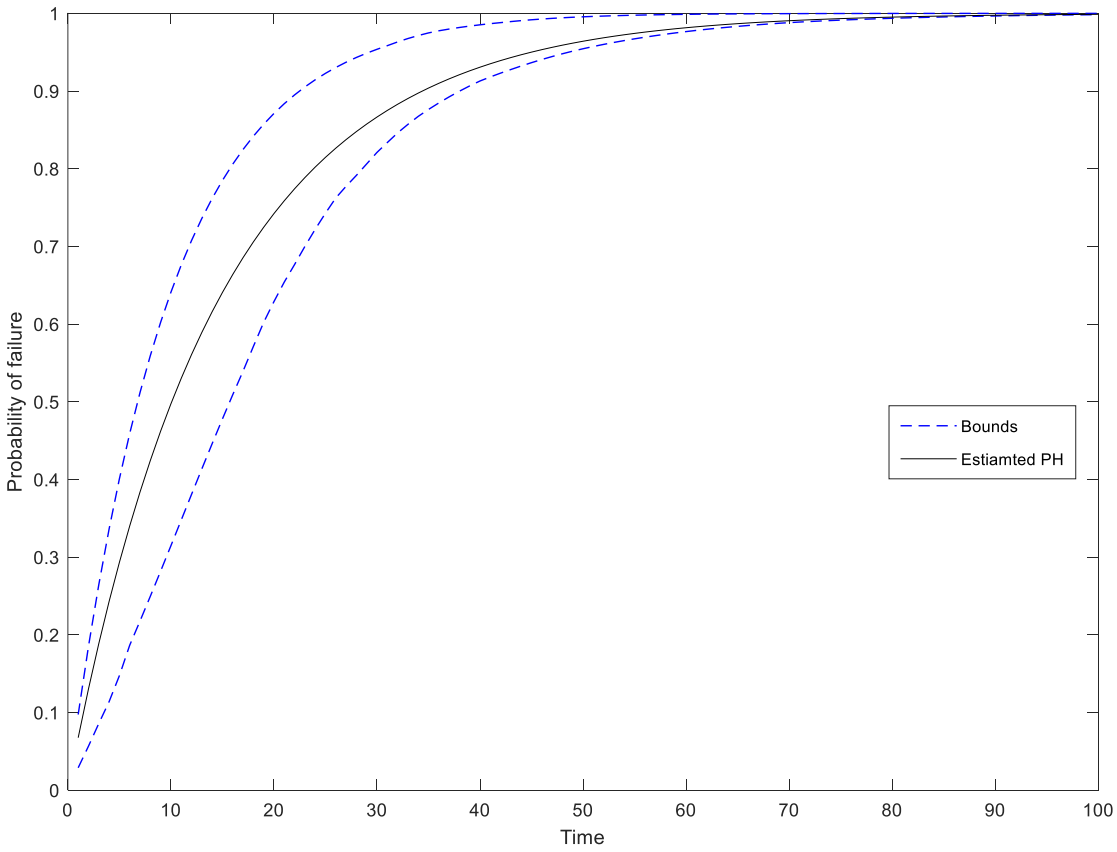

Figure 3.12: 90% bootstrap confidence interval of CDF based on the 3-phase Coxian

3.6 Conclusions and Future Work

Reliability estimation using aggregate data has been studied with only a few probability distributions. This work presents more flexible methods based on PH distributions to deal with such data for the first time. An EM algorithm is developed in this work by exploring the submatrices to utilize aggregate data. An alternative Bayesian method is also introduced to incorporate prior knowledge for parameter estimation. For the MLE method, model selection is performed through an MAP method. For the Bayesian method, model selection is concealed within the estimation procedure. Interval estimations are also obtained for the two methods. The flexibility of PH distribution for analyzing aggregate data with an arbitrary underlying distribution is explored in a simulation study and the capability of PH distribution is clearly illustrated. In addition, the proposed methods are

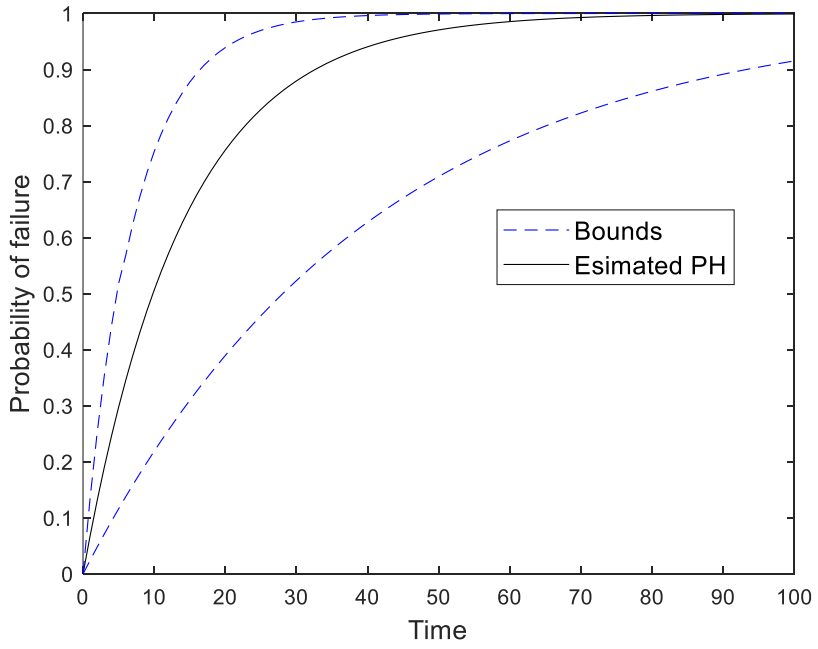

Figure 3.13: 90% Bayesian credible interval of CDF based on a 3-phase Coxian

successfully applied to the real dataset from RIAC. Considering that only a few probability distributions have been utilized for analyzing aggregate data, this work provides more flexible methods for analyzing aggregate failure-time data. Technically, the new EM algorithm, Fisher information and RJMCMC for PH distribution are used to analyze aggregate data for the first time.

For future work, interval estimation for PH distribution based on generalized pivotal quantity can be studied. Moreover, developing a nonparametric estimator based on aggregate data is a favorable while challenging research topic. Another interesting and common type of field data is time-censored aggregate data. The most common reason for collecting such data is to perform scheduled inspections. For time-censored aggregate data, each data point represents the number of failures in a certain period of time (e.g., during an inspection period). Unlike the aggregate data studied in this chapter, each time-censored aggregate time is not recorded at one of failures. The analysis of time-censored aggregate data has recently been discussed by Chen et al. (2020). Bayesian methods were provided for the Gamma, Inverse Gaussian, Weibull and Lognormal distributions. It is worth pointing out that the analysis of time-censored aggregate data through PH

distribution has not been discussed in the literature. The authors of this chapter have considered this research gap, and both ML and Bayesian estimation methods will be provided in their future work.

References

Asmussen S., Jensen J. L., & Rojas-Nandayapa L. (2016). On the Laplace transform of the lognormal distribution. *Methodology and Computing in Applied Probability, 18*(2), 441-58.

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting Phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, *23*(4), 419-441.

Ausín, M. C., Wiper, M. P., & Lillo, R. E. (2008). Bayesian prediction of the transient behaviour and busy period in short-and long-tailed GI/G/1 queueing systems. *Computational Statistics & Data Analysis, 52*(3), 1615-1635.

Beaulieu N. C., & Rajwani F. (2004). Highly accurate simple closed-form approximations to lognormal sum distributions and densities. *IEEE Communications Letters*, *8*(12), 709-11.

Beaulieu, N. C. & Xie, Q. (2004). An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology*, *53*(2), 479-489.

Bhaumik, D. K., Kapur, K., & Gibbons, R. D. (2009). Testing parameters of a gamma distribution for small samples. *Technometrics, 51*(3), 326-334.

Bladt, M., Esparza, L. J. R., & Nielsen, B. F. (2011). Fisher information and statistical inference for phase-type distributions. *Journal of Applied Probability, 48*(A), 277-293.

Bladt, M., Gonzalez, A., & Lauritzen, S. L. (2003). The estimation of Phase-type related functionals using Markov chain Monte Carlo methods. *Scandinavian Actuarial Journal, 2003*(4), 280-300.

Bučar, T., Nagode M., & Fajdiga M. (2004). Reliability approximation using finite Weibull mixture distributions. *Reliability Engineering & System Safety*, *84*(3), 241-51.

Buchholz, P., Kriege, J., & Felko, I. (2014). *Input Modeling with Phase-Type Distributions and Markov Models: Theory and Applications*. SpringerBriefs in Mathematics.

Chen, P., & Ye, Z. S. (2017a). Random effects models for aggregate lifetime data. *IEEE Transactions on Reliability, 66*(1), 76-83.

Chen, P. & Ye, Z. S. (2017b). Estimation of field reliability based on aggregate lifetime data. *Technometrics, 59*(1), 115-125.

Chen, P., Ye, Z. S., & Zhai, Q. (2020). Parametric analysis of time-censored aggregate lifetime data. *IISE Transactions, 52*(5), 516-527.

Cobb, B. R., Rumi, R., & Salmerŏn, A. (2012). Approximating the distribution of a sum of lognormal random variables. *Statistics and Computing*, *16*(3), 293-308.

Coit, D. W., & Dey, K. A. (1999). Analysis of grouped data from field-failure reporting systems. *Reliability Engineering & System Safety, 65*(2), 95-101.

Coit, D. W., & Jin, T. (2000). Gamma distribution parameter estimation for field reliability data with missing failure times. *IIE Transactions, 32*(12), 1161-1166.

Cui, L., & Wu, B. (2019). Extended Phase-type models for multistate competing risk systems. *Reliability Engineering & System Safety*, *181*, 1-16.

Delia, M. C., & Rafael, P. O. (2008). A maintenance model with failures and inspection following Markovian arrival processes and two repair modes. *European Journal of Operational Research , 186*(2), 694-707.

Denson, W., Crowell, W., Jaworski, P., & Mahar, D. (2014). *Electronic Parts Reliability Data 2014*. Reliability Information Analysis Center, Rome, NY, USA.

Elmahdy, E. E., & Aboutahoun, A. W. (2013). A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling. *Applied Mathematical Modelling*, *37*(4), 1800-1810.

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika, 82*(4), 711-732.

Horvath, A. v Telek, M. (2007). Matching more than three moments with acyclic phase type distributions. *Stochastic Models*, *23*(2), 167–194.

Jin, T., & Gonigunta, L. S. (2010). Exponential approximation to Weibull renewal with decreasing failure rate. *Journal of Statistical Computation and Simulation*, *80*(3), 273-285.

Karimi, S., Liao, H., & Pohl, E. (2019). A robust approach for estimating field reliability using aggregate failure time data. *Accepted in Annual Reliability and Maintainability Symposium (RAMS)*. Orlando: IEEE.

Kharoufeh, J. P., Solo, C. J., & Ulukus, M. Y. (2010). Semi-Markov models for degradation-based reliability. *IIE Transactions, 42*(8), 599-612.

Lam, C. L. J., & Le-Ngoc, T. (2007). Log-shifted gamma approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology, 56*(4), 2121-2129.

Li, J., Chen, J., & Zhang, X. (2019). Time-dependent reliability analysis of deteriorating structures based on phase-type distributions. *IEEE Transactions on Reliability*.

Liao, H. & Guo, H. (2013). A generic method for modeling accelerated life testing data. *Reliability and Maintainability Symposium (RAMS), Proceedings., Annual.* (1-6). Orlando, FL: IEEE.

Liao, H. & Karimi, S. (2017). Comparison study on general methods for modeling lifetime data with covariates. *Prognostics and System Health Management Conference (PHM-Harbin)* (1-5). Harbin: IEEE.

Mahar, D., Fields, W., Reade, J., Zarubin, P., & McCombie, S. (2011). *Nonelectronic Parts Reliability Data*. Reliability Information Analysis Center.

Marie, R. (1980). Calculating equilibrium probabilities for λ(n)/ck/1/n queues. *In Proceedings of the Performance 1980*, 117–125.

McGrory, C. A., Pettitt, A. N., & Faddy, M. J. (2009). A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis, 53*(12), 4311-4321.

Meeker, W. Q. & Escobar, L. A. (1995). Teaching about approximate confidence regions based on maximum likelihood estimation. *The American Statistician, 49*(1), 48-53.

Mehta, N. B., Wu, J., Molisch, A. F., & Zhang, J. (2007). Approximating a sum of random variables with a lognormal. *IEEE Transactions on Wireless Communications*, *6*(7), 2690-2699.

Okamura, H., Dohi, T., & Trivedi, K. S. (2011). A refined EM algorithm for PH distributions. *Performance Evaluation, 68*(10), 938-954.

Okamura, H., Watanabe, R., & Dohi, T. (2014). Variational Bayes for phase-type distribution. *Communications in Statistics-Simulation and Computation, 43*(8), 2031-2044.

OREDA (2009). *OREDA offshore Reliability Data Handbook*. Det Norske Veritas (DNV), Høvik, Norway.

Osogami, T., & Harchol-Balter, M. (2003). A closed-form solution for mapping general distributions to minimal PH distributions. *In International Conference on Modelling Techniques and Tools for Computer Performance Evaluation* 200-217, Springer, Berlin, Heidelberg.

Osogami, T., & Harchol-Balter, M. (2006). Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation*, *63*(6), 524-552.

Perreault, L. J., Thornton, M., Goodman, R., & Sheppard, J. W. (2015). A swarm-based approach to learning phase-type distributions for continuous time Bayesian networks. *In 2015 IEEE Symposium Series on Computational Intelligence*. 1860-1867. Cape Town: IEEE.

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 637-654.

Segovia, M. C., & Labeau, P. E. (2013). Reliability of a multi-state system subject to shocks using phase-type distributions. *Applied mathematical modelling, 37*(7), 4883-4904.

Starobinski, D., & Sidi, M. (2000). Modeling and analysis of power-tail distributions via classical teletraffic methods. *Queueing Systems, 36*(1-3), 243-267.

Telek, M., & Heindl, A. (2003). Matching moments for acyclic discrete and continuous phase-type distributions of second order,. *International Journal of Simulation*, *3*(3-4), 47-57.

Watanabe, R., Okamura, H., & Dohi, T. (2012). An efficient MCMC algorithm for continuous PH distributions. *Simulation Conference (WSC), Proceedings of the 2012 Winter* (1-12). Berlin: IEEE.

Xu, D., Xiao, X., & Haibo, Y. (2020). Reliability Evaluation of Smart Meters Under Degradation-Shock Loads Based on Phase-Type Distributions. *In IEEE Access*, *8*, 39734-39746.

Yamaguchi, Y., Okamura, H., & Dohi, T. (2010). A variational Bayesian approach for estimating parameters of a mixture of Erlang distribution. Communications in Statistics—Theory and Methods, *39*(13), 2333-2350.

4 A New Method for Analysis of Censored Aggregate Data Using Phase-type Distribution

Field failure-time data provide ample sources of valuable information for product reliability estimation. However, actual failure times of individual units are usually not reported in many situations. Instead, aggregate failure-time data that give cumulative failure times of multiple units are collected and sometimes implemented along with censoring. Analyzing such data raises big challenges to reliability estimation using existing statistical methods. So far, only a few probability distributions have been utilized to handle aggregate failure-time data while many widely used probability distributions are intractable. In this work, statistical methods using Phase-type (PH) distributions are proposed for analyzing censored aggregate failure-time data for the first time. Specially, a censored aggregate failure-time model based on the Coxian distribution is proposed, and an Expectation-Maximization (EM) algorithm for maximum likelihood (ML) estimation and a Bayesian alternative are developed for model parameter estimation. Moreover, the proposed statistical model is extended to handle data with covariates. A simulation study and two real-world examples are provided to illustrate the superior capability of the proposed method as opposed to the existing methods. Indeed, by mimicking the probability distributions that are the true underlying distributions while inapplicable of handling such data, the proposed methods provide practitioners with a collection of robust statistical tools to overcome this challenge.

4.1 Introduction

4.1.1 Background and motivation

Field failure-time data provide ample sources of valuable information for product reliability estimation as they are collected under real use conditions. Therefore, a number of organizations, such as the U.S. Department of Defense, have collected a large volume of field failure data (Denson et al., 2014; Mahar et al., 2011; OREDA, 2009). However, due to the less controlled data collection processes, field data are in various formats and often have missing values. Specially, data aggregation happens when cumulative or combined data is provided instead of the failure times

of individual units. A type of aggregated data frequently seen is type-I censored or time-censored aggregate data. Such data is reported as the cumulative number of units used up to the end of an operating time. This type of data is referred to as *censored aggregate failure-time data* and is the subject of this work.

One way to report a censored aggregate data point is (t, n) that contains a time value t and an integer value n, where t is the cumulative operating time of $n + 1$ components with n failures. For example, in a certain component position, a component is replaced with a new one immediately when it fails, and the number of actual failures (i.e., n failures plus one working component) at the end of a specific period of time t is reported. In many applications, it is common that t is equal for all the censored aggregate data points. This type of data may be collected as the results of scheduled inspections or regular reports at the end of a certain time period. However, in general, the values of t may be different for different data points. In this work, this data format without assuming a common value of t for different data points is adopted. Note that, this type of data is also called *count data* in queueing models and many other applications. Indeed, a component's lifetime is equivalent to an inter-arrival time in count data, and each failure is an event.

To model count data, the most common method is to assume a Poisson process, and when covariates are involved, a Poisson regression model is often considered. However, in many situations such Poisson process models may not adequately describe count data. For example, the equidispersion assumption is often violated, and the memoryless property of the Exponential distribution is not valid for a failure caused by damage accumulation due to corrosion, fatigue, fracture, etc. (Coit & Jin, 2000). It is worth pointing out that aggregate failure-time data must be modeled based on the convolution of the underlying failure-time distributions. So far, only the Exponential, Normal, Gamma and Inverse Gaussian (IG) distributions have been used because their closed-form expressions for aggregate data are available. However, for other widely used probability distributions, such as Weibull, Lognormal and Extreme Value distributions, their closed-form expressions for such data are not attainable. Clearly, it is valuable to unleash our potential in analyzing aggregate failure-time data without being limited by a few probability distributions.

Phase-type (PH) distributions are a family of probability distributions that are quite flexible in mimicking the distributions of nonnegative random variables. In this work, PH distributions are utilized for analyzing censored aggregate failure-time data for the first time. A maximum likelihood (ML) estimation method and a Bayesian alternative are developed for modeling such data without and with covariates. The goal is to assist practitioners in using such abundant field data for reliability estimation without relying on some prior knowledge about the underlying failure-time distributions. On the other hand, this work provides a tractable way to overcome the challenge when the true underlying distribution does not have a closed-form expression for such data.

4.1.2 Related work

To represent the event and inter-arrival time relationship, the Exponential distribution probably is the most widely used distribution. Coit and Dey (1999) tested the Exponential distribution assumption and warned against using it in some cases. Winkelmann (1995) used a Gamma count probability model, where the inter-arrival times follow the Gamma distribution. McShane et al. (2008) studied a count process with Weibull inter-arrival times. They derived a Weibull count model using a polynomial expansion for closed-form inference. Kharrat et al. (2019) developed an R package (i.e., Countr) that provides an accessible way for performing regression on count data based on renewal processes. Unlike those models that assume the independence of individual lifetimes, several models, called occurrence-dependent models, were utilized for cases where each event depends on the number of prior events (Chen et al., 2015; Yin et al., 2016). Recently, Xiao et al. (2020) proposed a nonparametric Bayesian modeling and estimation method for renewal processes.

Regarding aggregate failure-time data analysis, other than the Exponential distribution, Gamma (Coit & Jin, 2000) and IG distributions (Chen & Ye, 2017) have been successfully implemented. Specially, the analysis of censored aggregate failure-time data was investigated by Chen et al. (2020). They provided MLE methods for Gamma and IG distributions as well as an approximate Bayesian computation algorithm for the Lognormal and Weibull distributions.

In many applications, the true distributions may not be directly used due to their intractability. To overcome such technical challenges, the lognormal distribution (Beaulieu & Rajwani, 2004; Beaulieu & Xie, 2004; Lam & Le-Ngoc, 2007; Mehta et al., 2007; OREDA, 2009), mixture of Weibull (Bučar et al., 2004; Elmahdy & Aboutahoun, 2013; Jin & Gonigunta, 2010), and Laplace method (Asmussen et al., 2016; Rizopoulos et al., 2009; Rue et al., 2009) have been widely used in distribution approximation. To take the advantage of PH distributions in mimicking other probability distributions, a large bulk of work on distribution approximation is dedicated to the use of PH distributions. Osogami and Harchol-Balter (2003), Osogami and Harchol-Balter (2006) and Horváth and Telek (2007) used moment matching methods to approximate a general distribution with a PH distribution. Matching the shape of distributions with PH distributions has also been considered (Riska et al., 2004; Starobinski & Sidi, 2000).

Regarding parameter estimation for PH distributions, Asmussen et al. (1996) developed an expectation-maximization (EM) algorithm and proposed to minimize the information divergence in density approximation. To reduce the high computational complexity of the EM algorithm for PH distributions, several fast estimation methods have been developed (Okamura et al., 2014; OREDA, 2009; Yamaguchi et al., 2010). Since some information is lost or approximated in such fast algorithms, the reduction in estimation accuracy and precision is unavoidable, and the most proper method should be selected according to the problem. In addition to the ML estimation method, the Bayesian alternative provides another direction for parameter estimation of PH distributions. Specially, Markov Chain Monte Carlo methods are usually implemented (Ausıén et al., 2008; McGrory et al., 2009; Watanabe et al., 2012).

Although PH distributions have been widely used for modeling individual failure-time data, the work of Karimi et al. (2020) is the only exception that used PH distributions for aggregate failure-time data. However, the statistical estimation methods proposed in their work are only valid for analyzing aggregate failure-time data without censoring. Moreover, in many real world applications, the data are often collected with some covariates. Specially, count regression models that account for the effects of socio-economics, health insurance coverage and disease status

have been applied vastly to healthcare problems (Cameron & Johansson, 1997). In the area of reliability, many static and dynamic covariates are recorded for engineering systems (Liu & Pan, 2020; Meeker & Hong, 2014). To handle censored aggregate failure-time data with covariates, a PH-based regression model is also developed in this work. Note that the work by Chen et al. (2020) on the analysis of censored aggregate failure-time data was focused on a couple of specific probability distributions without considering covariates. In this work, PH distributions are utilized as a flexible tool for the analysis of such data with covariates. It is also worth pointing out that our proposed PH-based models and statistical estimation methods can be implemented in other applications, such as modeling the length-of-stay of patients (Gu et al., 2019), stochastic operating room scheduling (Varmazyar et al., 2020), and warranty data analysis (He et al., 2018).

4.1.3 Overview

The remainder of this chapter is organized as follows. Section 4.2 provides the preliminaries on PH distributions and the key definitions. Section 4.3 introduces the basic model for censored aggregate data. The maximum likelihood estimation method and the Bayesian alternative for cases without and with covariates are described in Section 3.4 and Section 3.5, respectively. Numerical examples including a simulation study and studies on two real datasets are provided in Section 3.6. Finally, Section 3.7 draws conclusions.

4.2 Preliminaries on PH Distributions

A continuous PH distribution, which is used in this work due to the type of data under study, describes the time to absorption of a continuous-time Markov chain (CTMC). Consider $X(t)_{t\geq 0}^{\infty}$ as an absorbing Markov process with N transient states $1, 2, \cdots, N$ and one absorbing state $N+1$. Let $\mathbb{Q}$ be the infinitesimal generator of $X(t)_{t\geq 0}^{\infty}$ given by:

$$\mathbb{Q} = \begin{pmatrix} \mathbb{S} & \mathbf{S}^0 \\ \mathbf{0}' & 0 \end{pmatrix}, \tag{4.1}$$

where $\mathbf{0}' = [0, 0, \cdots, 0]$, $\mathbb{S}$ is the transition rate matrix for the transient states, and $\mathbf{S}^0 = -\mathbb{S}\mathbf{1}$ represents the absorption rates with $\mathbf{1}' = [1, 1, \cdots, 1]$ (Buchholz et al., 2014). Then, the probability density function (pdf) $f(\cdot)$ and the cumulative distribution function (CDF) $F(\cdot)$ of time to absorption are:

$$f(t) = \boldsymbol{\pi} e^{\mathbb{S}t}\mathbf{S}^0, \qquad F(t) = 1 - \boldsymbol{\pi} e^{\mathbb{S}t}\mathbf{1}, \tag{4.2}$$

respectively, where $\pi = [\pi(1), \pi(2), \cdots, \pi(N)]$ is the vector of initial probabilities with π_i being the probability that the process starts at the i^{th} state. Throughout the chapter, it is assumed that the probability that the process starts from the absorbing state is zero.

The probability distribution defined by equation (4.2) for the CTMC $X(t)_{t\geq 0}^{\infty}$ is called an N-phase PH distribution. The following matrix shows a general acyclic PH distribution transition rate matrix:

$$\mathbb{S} = \begin{pmatrix} -\lambda_1 & p_{12}\lambda_1 & p_{13}\lambda_1 & \cdots & p_{1N}\lambda_1 \\ 0 & -\lambda_2 & p_{23}\lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & -\lambda_{N-1} & p_{(N-1)N}\lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\lambda_N \end{pmatrix} \tag{4.3}$$

where $0 \leq p_{ij} \leq 1$, $\sum_{j=1}^{N} p_{ij} \leq 1$, $i = 1, 2, \cdots, N-1$, $j = 1, 2, ..., N$ and $i < j$. In practice, since a product wears out until failure, the acyclic form of the distribution is a reasonable choice.

There are a number of special forms of PH distributions which have much sparser transition matrices than the general PH distribution. Some useful special cases are Exponential, Erlang, Hyper-exponential, Hypo-exponential, Hyper-Erlang, and Coxian distributions. Specially, the N-phase Coxian distribution has the following transition rate matrix and vector of initial probabilities:

$$\mathbb{S} = \begin{pmatrix} -\lambda_1 & p_1\lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & p_2\lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -\lambda_{N-1} & p_{N-1}\lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\lambda_N \end{pmatrix}, \boldsymbol{\pi} = \begin{bmatrix} 1 & 0 & \cdots & 0 \end{bmatrix}_{1\times N}. \tag{4.4}$$

Figure 4.1 shows a schematic of the N-phase Coxian distribution. The corresponding CTMC always starts in phase 1, and it can jump either to the immediately next state or to the absorbing state. Note that the order of the number of parameters of a general acyclic PH distribution is $O(N^2)$, while for the Coxian distribution it is $O(N)$. Generally, the Coxian distribution has $2N-1$ parameters. It is worth pointing out that although the Coxian distribution has a simple structure, it can uniquely represent any general acyclic PH distribution.

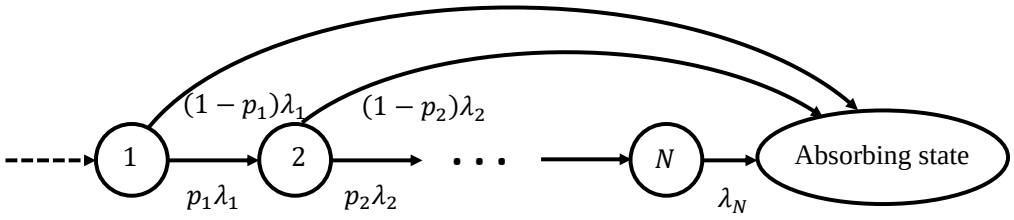

Figure 4.1: CTMC for N-Phase Coxian distribution.

Although the statistical estimation methods provided in this chapter can be used for general PH distributions, to increase the computational efficiency, we will use the Coxian distribution in the

rest of our study. Specially, we consider the following reparameterization:

$$\mathbb{S} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -(\lambda_{N-1}+\mu_{N-1}) & \lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\mu_N \end{pmatrix}. \tag{4.5}$$

Then, the absorption rate matrix can be expressed as $\mathbf{S}^0 = [\mu_1, \mu_2, ..., \mu_N]^T$.

4.3 PH Model for Censored Aggregate Data

We first provide a general framework for PH distributions in relation to censored aggregate failure-time data. We consider a component in a certain component position of the system. As the first unit fails, it is immediately replaced by a new identical unit. Let t be a certain period of time, n be the number of replacements during t, and τ_l be the lifetime of unit l (i.e., the lth interarrival time), $l = 1, 2, \cdots, n+1$, which is unknown. Then, the probabilistic model for a censored aggregate failure-time data point (t, n) without covariates can be expressed as:

$$P_r(\sum_{l=1}^{n+1} \tau_l > t, \sum_{l=1}^{n} \tau_l < t) = \int_0^t R(t-y) f_n(y) dy, \tag{4.6}$$

where $f_n(\cdot)$ represents the pdf of the sum of n component lifetimes (time of the nth arrival), and $R(\cdot) = 1 - F(\cdot)$ is the reliability function of each component. Clearly, the integration in equation (4.6) can be very intensive to do for a PH distribution. However, by looking at this probability from the viewpoint of the PH distribution structure, we can derive the exact answer. To this end, we first provide three key remarks.

Remark 1. The transition matrix of a PH distribution: From the infinitesimal generator of an

absorbing Markov process given in equation (4.1), the transition matrix $\boldsymbol{P}_t = e^{\mathbb{Q}t}$ is defined as:

$$\boldsymbol{P}_t = e^{\mathbb{Q}t} = \begin{pmatrix} e^{\mathbb{S}t} & \mathbf{1} - e^{\mathbb{S}t}\mathbf{1} \\ \mathbf{0} & 1 \end{pmatrix}, \tag{4.7}$$

where $\boldsymbol{P}_t(i,j)$ represents the probability of the process being in phase j at time t given that the Markov process has started at phase i (Buchholz et al., 2014). In other words, considering the Markov process $X(t)_{t\geq 0}^{\infty}$, we have $\boldsymbol{P}_t(i,j) = P_r(X(t) = j | X(0) = i)$. P_r notation is used to represent probability.

Remark 2. The convolution of a PH distribution: Assuming random variables A and B follow PH distributions with transition rate matrices $\mathbb{S}^{(A)}$ and $\mathbb{S}^{(B)}$, respectively, the variable $C = A + B$ follows a PH distribution with transition rate matrix:

$$\mathbb{S}^{(C)} = \begin{pmatrix} \mathbb{S}^{(A)} & \mathbf{S}^0\boldsymbol{\pi}^{(B)} \\ \mathbf{0} & \mathbb{S}^{(B)} \end{pmatrix}. \tag{4.8}$$

The corresponding initial probability vectors $\boldsymbol{\pi}^{(A)}$ and $\boldsymbol{\pi}^{(B)}$ will make $\boldsymbol{\pi}^{(C)} = [\boldsymbol{\pi}^{(A)}, \mathbf{0}_{1\times N_B}]$, where N_B is the number of phases of the distribution of B. Deductively, the convolution of $n+1$ independent and identically distributed N-phase PH variables with transition rate matrix $\mathbb{S}_{N\times N}$ and initial probability vector has: $\boldsymbol{\pi}_{1\times N}$,

$$\mathbb{S}^{conv} = \begin{pmatrix} \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \mathbf{0} \\ \vdots & & \ddots & \mathbb{S} & \mathbf{S}^0\boldsymbol{\pi} \\ \mathbf{0} & \cdots & \cdots & \mathbf{0} & \mathbb{S} \end{pmatrix}_{N(n+1)\times N(n+1)}, \boldsymbol{\pi}^{conv} = [\boldsymbol{\pi}_{\mathbf{1}\times \boldsymbol{N}}, \mathbf{0}_{\mathbf{1}\times \boldsymbol{Nn}}]. \tag{4.9}$$

Remark 3. For a censored aggregate failure-time data point (t, n), we know that t is a point in time between the n^{th} failure and the $n+1^{\text{th}}$ failure. In the process with the $n+1$ components, each

of which has N phases, the infinitesimal generator and the transition matrix can be expressed as:

$$\mathbb{Q}^{conv} = \left(\begin{array}{c|c} \mathbb{S}^{conv} & -\mathbb{S}^{conv}\mathbf{1} \\ \hline \mathbf{0}' & 0 \end{array}\right), \tag{4.10}$$

and

$$\boldsymbol{P}_t = e^{\mathbb{Q}^{conv}t} = \left(\begin{array}{c|c} e^{\mathbb{S}^{conv}} & \mathbf{1} - e^{\mathbb{S}^{conv}}\mathbf{1} \\ \hline \mathbf{1}' & 1 \end{array}\right). \tag{4.11}$$

In summary, the aggregate failure time of $n+1$ identical units follows a $N(n+1)$-phase PH distribution with a transition rate matrix as stated in Remark 2. At time t, n components have failed, and the process lies in somewhere in its last N phases (i.e., between $nN+1$ and $N(n+1)$). Moreover, we know that the process has started in the first N phases, making the square matrix $\boldsymbol{P}_t^{\star}$, as in Figure 4.2, have all the probabilities that we need. Hence, the closed-form expression for equation (4.6) is:

$$P_r(\sum_{l=1}^{n+1}\tau_l > t, \sum_{l=1}^{n}\tau_l < t) = \boldsymbol{\pi}^{conv}e^{\mathbb{S}^{conv}t}\boldsymbol{e} = \boldsymbol{\pi}\boldsymbol{P}_t^{\star}\mathbf{1}, \tag{4.12}$$

where $\boldsymbol{\pi}^{conv}$ is the initial probability vector for the specific data point, $\boldsymbol{\pi}$ is the initial probability vector of a single component, $\mathbf{1}$ is a vector of ones, and $\boldsymbol{e}$ is a vector of all zeros except with ones in $(Nn+1)$th to $N(n+1)$th positions. It is noteworthy that if $\boldsymbol{\pi}^{conv}e^{\mathbb{S}^{conv}t}\boldsymbol{e}$ is used, the calculation will be much less efficient than using $\boldsymbol{\pi}\boldsymbol{P}_t^{\star}\mathbf{1}$. Indeed, using $\boldsymbol{\pi}\boldsymbol{P}_t^{\star}\mathbf{1}$ will decrease the computation effort by $O(1/n^2)$ due to the smaller size of the adopted matrices.

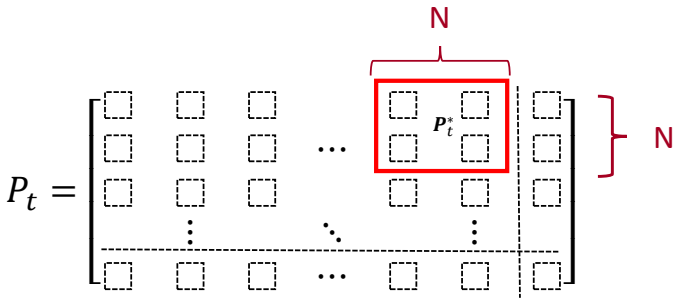

Figure 4.2: The transition matrix for data point (t, n) with the marked square matrix $\boldsymbol{P_t^\star}$.

In the presence of covariates, each data point can be expressed in the form of $(t, \boldsymbol{X}, n)$, where $\boldsymbol{X}$ is the vector of covariates that affect the component's lifetime. For illustration, we consider an accelerated failure time model to incorporate the life-stress relationship such that the corresponding transition matrix can be expressed as $\boldsymbol{P}_{t,X} = e^{\exp(\boldsymbol{b}^T \boldsymbol{X})\mathbb{Q}^{conv} t}$, where $\boldsymbol{b}$ is the regression coefficient vector. Similar to equation (4.12), the censored aggregate failure-time model for data $(t, \boldsymbol{X}, n)$ with covariates can be expressed as:

$$P_r(\sum_{l=1}^{n+1} \tau_l > t, \sum_{l=1}^{n} \tau_l < t|\boldsymbol{X}) = \boldsymbol{\pi}^{conv} e^{\exp(\boldsymbol{b}^T \boldsymbol{X})\mathbb{S}^{conv} t} \boldsymbol{e} = \boldsymbol{\pi} \boldsymbol{P}_{t,X}^{\star} \mathbf{1}, \tag{4.13}$$

where the position of the corresponding submatrix $\boldsymbol{P}_{t,X}^{*}$ is the same as that of $\boldsymbol{P}_t^*$ indicated in Figure 4.2.

4.4 Maximum Likelihood Estimation Method

The likelihood function for the failure times of m individual components can be expressed as:

$$\mathcal{L}((\boldsymbol{\pi}, \mathbb{S})|\boldsymbol{\tau}) = \prod_{k=1}^{m} \boldsymbol{\pi} e^{\mathbb{S} t_k} \mathbf{S^0}, \tag{4.14}$$

where $\boldsymbol{\tau} = [t_1, t_2, \cdots, t_m]$ is the vector containing the individual failure times, and $\boldsymbol{\pi}$ is the initial probability vector of a single component that is assumed to be the same for all the units in the rest of this chapter. Because of the presence of matrix exponential in the equation, it is difficult to

directly maximize the likelihood function for finding the ML estimates of the model parameters.

Asmussen et al. (1996) proposed an EM algorithm for estimating the parameters of general PH distributions, which can handle individual failure times. As each data point provides only the time of absorption and no information about the transitions between phases, the data can be regarded as having missing values. In this algorithm, the number of times the process has started in each phase ($B_.$), the sojourn time of each phase ($Z_.$) and the number of transitions between phases ($N_.$) are estimated from the data. The likelihood function considering the complete information $\boldsymbol{z}$ including the times of arriving at each phase and the sojourn times is:

$$\mathcal{L}((\boldsymbol{\pi}, \mathbb{S})|\boldsymbol{\tau}) = f(\boldsymbol{z}|(\boldsymbol{\pi}, \mathbb{S})) = \prod_{i=1}^{N} \pi(i)^{B_i} \prod_{i=1}^{N} e^{Z_i \mathbb{S}(i,i)} \prod_{i=1}^{N} \prod_{j=1}^{N+1} \mathbb{S}(i,j)^{N_{ij}}. \tag{4.15}$$

The algorithm consists of two steps performed iteratively. In the expectation step (E-step), the unobserved variables are estimated based on the newest estimates of the parameters. In the maximization step (M-step), the parameters are re-estimated based on the observed and estimated data.

It is worth pointing out that this EM algorithm is not applicable to censor aggregate failure-time data. In this section, two EM algorithms will be developed to handle censored aggregate failure-time data without and with covariates. The two algorithms are applicable to general PH distributions, and can be simply applied to the Coxian distribution defined by equation (4.5) and the vector of initial probabilities $\boldsymbol{\pi} = [1, 0, \cdots, 0]_{1\times N}$.

4.4.1 Case without covariates

4.4.1.1 Likelihood function

For censored aggregate failure-time data $(\boldsymbol{t}, \boldsymbol{n}) = [(t_1, n_1), (t_2, n_2), \cdots, (t_m, n_m)]$, the likelihood function is:

$$\mathcal{L}((\boldsymbol{\pi}, \mathbb{S})|(\boldsymbol{t}, \boldsymbol{n})) = \prod_{k=1}^{m} \boldsymbol{\pi}_k \boldsymbol{P}_{t_k} \boldsymbol{e}_k = \prod_{k=1}^{m} \boldsymbol{\pi} \boldsymbol{P}^{\star}_{t_k} \mathbf{1}, \tag{4.16}$$

where $\boldsymbol{\pi}_k$ is a $1 \times N(n_k+1)$ vector as given in equation (4.9), and $\boldsymbol{e}_k$ is a vector of zeros except with ones in (Nn_k+1)th to $N(n_k+1)$th positions.

4.4.1.2 EM algorithm for data without covariates In the context of censored aggregate failure-time data, the distribution parameters are different for different data points. Specifically, $\boldsymbol{\pi}_k$ and $\mathbb{S}_k$ are determined by equation (4.9) and $\mathbf{S}_k^0 = -\mathbb{S}_k\mathbf{1}$. We first define the unobserved variables. For better presentation, we use i' and j' for any of $i+N(l-1)$ with $l = 1, 2, \cdots, n+1$ and $i = 1, 2, \cdots, N$. Specially, we define:

- $B_i = \sum_{k=1}^{m} \mathbf{1}_{J_0^{[k]}=i'}$: the number of times the Markov process started in state $i+N(l-1)$, $i = 1, 2, \cdots, N, l = 1, 2, \cdots, n+1$.

- $Z_i = \sum_{k=1}^{m} \int_0^{\infty} \mathbf{1}_{J_v^{[k]}=i'} dv$: the length of time the Markov process spent in state $i+N(l-1)$, $i = 1, 2, \cdots, N, l = 1, 2, \cdots, n+1$.

- $N_{ij} = \sum_{k=1}^{m} \sum_{v=0}^{\infty} \mathbf{1}_{J_{v\epsilon}^{[k]}=i', J_{(v+1)\epsilon=j'}^{[k]}}$: the number of times the Markov process jumped from state $i+N(l-1)$, to $j+N(l-1)$, $i, j = 0, 1, 2, \cdots, N$, $l = 1, 2, \cdots, n+1$, and $J_v^{[k]}$ shows the phase (state) of the process at the v^{th} time interval for component k.

Then, the E-step formulas for each of the censored aggregate failure-time data are as follows (see Appendix for the detailed derivations):

$$\begin{aligned} E[B_i|T=t] = \frac{1}{n+1}\Big(&\frac{\boldsymbol{\pi}(i)\boldsymbol{e}_i' e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}} \\ &+ \frac{\int_0^t \sum_{v=1}^{n} \boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{(v-1)N+1:vN}\boldsymbol{S}_{v-1}\boldsymbol{\pi}(i)e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N}du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}\Big), \end{aligned} \tag{4.17}$$

$$E[Z_i|T=t] = \frac{1}{n+1}\frac{\int_0^t \sum_{v=1}^{n+1} \boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{i+N(v-1)}\boldsymbol{e}_{i+N(v-1)}' e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N}du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}, \tag{4.18}$$

$$E[N_{ij}|T=t] = \frac{1}{n+1} \times \frac{\int_0^t \sum_{v=1}^{n+1} (\boldsymbol{\pi} e^{\mathbb{S}u} \boldsymbol{e}_{i+(N(v-1))})(\mathbb{S}_{ij})(\boldsymbol{e}'_{j+N(v-1)} e^{\mathbb{S}(t-u)} \boldsymbol{e}_{nN+1:(n+1)N}) du}{\boldsymbol{\pi} e^{\mathbb{S}t} \boldsymbol{e}_{nN+1:(n+1)N}}, \tag{4.19}$$

$$E[N_{i0}|T=t] = \frac{1}{n+1} \frac{\int_0^t \sum_{v=1}^{n} (\boldsymbol{\pi} e^{\mathbb{S}u} \boldsymbol{e}_{i+(N(v-1))})(\mathbf{S}_i)(\boldsymbol{\pi}_{v+1} e^{\mathbb{S}(t-u)} \boldsymbol{e}_{nN+1:(n+1)N}) du}{\boldsymbol{\pi} e^{\mathbb{S}t} \boldsymbol{e}_{nN+1:(n+1)N}}, \tag{4.20}$$

where e_i is a vector with 1 in the ith entry and 0 elsewhere. Similarly, $e_{i:j}$ has 1 in entries $i, i+1, \cdots, j$ and 0 elsewhere. After performing the E-step, the following M-step formulas:

$$\begin{aligned} \hat{\boldsymbol{\pi}}(i) &= E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[B_i], & \hat{\mathbb{S}}(i,j) &= \frac{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{ij}]}{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i]}, \\ \hat{\mathbf{S}}^0(i) &= \frac{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[N_{in+1}]}{E_{(\boldsymbol{\pi},\mathbb{S}),\boldsymbol{\tau}}[Z_i]}, & \hat{\mathbb{S}}(i,i) &= -(\hat{\mathbf{S}}^0(i) + \sum_{i \neq j}^{n} \hat{\mathbb{S}}(i,j)), \end{aligned} \tag{4.21}$$

can be applied to update the parameter estimates by maximizing the likelihood function. Note that for the Coxian distribution, $\boldsymbol{\pi}$ is given as $\boldsymbol{\pi} = [1, 0, \cdots, 0]_{1\times N}$, so there is no need to calculate B_i and $\hat{\pi}(i)$, $i = 1, 2, \cdots, N$.

Algorithm 1 provides the detailed steps of the proposed EM algorithm. Note that the result of the EM algorithm is monotonically improving when performing the E-step and M-step iteratively until the convergence criterion is met (i.e., the difference between the likelihood values in the last two iterations is less than a predetermined threshold ϵ).

Algorithm 1: EM algorithm for fitting a PH distribution without covariates

Result: MLE of parameters in N-phase PH distribution

Data: $(\boldsymbol{t}, \boldsymbol{n})_{m,:}$

Require: Initial estimates of the model parameters $\boldsymbol{\pi}$, $\mathbb{S}$, constant ϵ for convergence criterion, Previous Likelihood $\leftarrow -\infty$, Current Likelihood $\leftarrow \infty$.

while $\left|\frac{\textit{Previous Likelihood}-\textit{Current Likelihood}}{\textit{Previous Likelihood}}\right| > \epsilon$ **do**

  $\boldsymbol{\pi}_k \leftarrow ((n_k+1)$th convolution of $\boldsymbol{\pi})$, $\mathbb{S}_k \leftarrow ((n_k+1)$th convolution of $\mathbb{S})$ for $k = 1, 2, \cdots, m$;

  Calculate B_i, Z_i, N_{ij}, N_{iN+1} for $i, j = 1, 2, \cdots, N$ and $i \neq j$, using equations (4.17) – (4.20);

  Update the estimates of distribution parameters $\boldsymbol{\pi}$, $\mathbb{S}$ using equation set (4.21);

  Previous Likelihood $\leftarrow$ Current Likelihood;

  Current Likelihood $\leftarrow \mathcal{L}((\boldsymbol{\pi}, \mathbb{S})|(\boldsymbol{t}, \boldsymbol{n}))$;

end

4.4.2 Case with covariates

4.4.2.1 Likelihood function For censor aggregate failure-time data with covariates $(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n}) = [(t_1, \boldsymbol{X}_1, n_1), \cdots, (t_m, \boldsymbol{X}_m, n_m)]$, where the first entry of $\boldsymbol{X}_k$ is 1 to include the intercept in the vector of coefficients. When using the log-linear model as the life-stress relationship, the new parameter matrix of the model becomes:

$$\mathbb{S}_k^* = \exp(\boldsymbol{b}^T \boldsymbol{X}_k)\mathbb{S}, \qquad \mathbf{S}_k^{0*} = \mathbb{S}_k^* \mathbf{1} = \exp(\boldsymbol{b}^T \boldsymbol{X}_k)\mathbf{S^0}. \tag{4.22}$$

Then, the corresponding likelihood function can be expressed as:

$$\mathcal{L}((\boldsymbol{\pi}, \mathbb{S}, \boldsymbol{b})|(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n})) = \prod_{k=1}^{m} \boldsymbol{\pi} \boldsymbol{P}^{\star}_{t_k, X_k} \mathbf{1}, \tag{4.23}$$

where $\boldsymbol{P}^{\star}_{t_k,X_k} = (e^{\mathbb{S}^*_k t_k})_{1:N,N(n_k)+1:N(n_k+1)}$ is the submatrix of the transition matrix corresponding to the kth censored aggregate failure-time data point.

4.4.2.2 EM algorithm for data with covariates

Algorithm 2 gives the detailed steps of the proposed EM algorithm.

Algorithm 2: EM algorithm for fitting a PH distribution with covariates

Result: MLE of parameters in N-phase PH distribution with covariates

Data: $(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n})_{m,:}$

Require: Initial estimates of model parameters $\boldsymbol{\pi}$, $\mathbb{S}$ and $\boldsymbol{b}$, constant ϵ for convergence criterion, Previous Likelihood $\leftarrow -\infty$, Current Likelihood $\leftarrow \infty$, $t^*_k \leftarrow t_k \exp(\boldsymbol{b}^T X_k)$ for $k = 1, 2, \cdots, m$.

while $\left|\frac{\textit{Previous Likelihood}-\textit{Current Likelihood}}{\textit{Previous Likelihood}}\right| > \epsilon$ **do**

- $\boldsymbol{\pi}_k \leftarrow ((n_k+1)$th convolution of $\boldsymbol{\pi})$, $\mathbb{S}_k \leftarrow ((n_k+1)$th convolution of $\mathbb{S})$ for $k = 1, 2, \cdots, m$;
- Calculate B_i, Z_i, N_{ij}, N_{iN+1} for $i, j = 1, 2, \cdots, N$ and $i \neq j$, using equations (4.17) – (4.20). Replace $\boldsymbol{t}$ by $\boldsymbol{t}^*$ in all equations;
- Update the estimates of distribution parameters $\boldsymbol{\pi}$, $\mathbb{S}$ using equation set (4.21);
- Update $\boldsymbol{b}$ using random search;
- $t^*_k \leftarrow \exp(\boldsymbol{b}^T \boldsymbol{X}_k) t_k$ for $k = 1, 2, \cdots, m$;
- Previous Likelihood $\leftarrow$ Current Likelihood;
- Current Likelihood $\leftarrow \mathcal{L}((\boldsymbol{\pi}, \mathbb{S}, \boldsymbol{b})|(\boldsymbol{t}^*, \boldsymbol{X}, \boldsymbol{n}))$;

end

In each iteration, matrix $\mathbb{S}$ is estimated just as the case without covariates. Afterwards, the coefficient vector $\boldsymbol{b}$ is updated through a random search algorithm. Specially, the search is performed by maximizing the log-likelihood value in an interval $b_u \pm \gamma$ for each coefficient b_u in $\boldsymbol{b}$. The hyperparameter γ can be defined based on the requirements of the problem. The value of $\boldsymbol{b}$ that gives the best likelihood value of equation (4.23) is kept as the current estimate of the coefficient

vector. The reason that a gradient-based algorithm cannot be used to find the optimal value of the coefficients lies in the multimodality of the likelihood function. Subsequently, $t_k^* = \exp(\boldsymbol{b}^T \boldsymbol{X}_k) t_k$ is updated, as if we get a corrected set of data. The algorithm stops when the convergence criterion is met (i.e., the difference between the likelihood values in the last two iterations is less than a predetermined threshold ϵ).

It is worth pointing out that some of the steps in Algorithm 2 are not different from those for the case without covariates. By looking at the likelihood function in equation (4.23), we can see the life-stress relationship can be interpreted as time-scaling. In other words, under $\boldsymbol{X}_k = 0$ we have standard time values, t_k. However, if X_k is nonzero, we can see that the time value changes to $t_k^* = \exp(\boldsymbol{b}^T \boldsymbol{X}_k) t_k$, called the standardized time value. Therefore, by scaling all the time values, we can use t_k^* in the calculations as the regular time value. So, by updating $\boldsymbol{b}$ in each iteration we are, in fact, updating the standardized data and getting closer and closer to the true scale.

4.5 Bayesian Estimation Method

a Bayesian alternative is also provided in this work for fitting a PH distribution to censored aggregate failure-time data without and with covariates. Note that McGrory et al. (2009) developed a Bayesian method based on a Reversible Jump Markov Chain Monte Carlo (RJMCMC) for fitting Coxian distributions. However, their method cannot be directly applied to the censored aggregate failure-time data. In this section, we will develop the Bayesian method specially for the Coxian distribution by modifying the RJMCMC method. One superiority of the RJMCMC method in comparison to many other available Bayesian methods, such as Gibbs sampling, is its automatic model selection capability.

4.5.1 The proposed Bayesian model

4.5.1.1 Case without covariates For analyzing censored aggregate failure-time data, parameter updates can be done with regard to one component lifetime, similar to Section 4.4. For a Coxian distribution, the parameters are divided into three groups of N, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$, where N is the

number of phases of Coxian that is not predetermined in this model, and the parameter vectors of $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_{N-1}]$ and $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_N]$ are shown in the following single component transition rate matrix:

$$\mathbb{S} = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & \ddots & -(\lambda_{N-1}+\mu_{N-1}) & \lambda_{N-1} \\ 0 & \cdots & \cdots & 0 & -\mu_N \end{pmatrix} \tag{4.24}$$

In this work, Gamma distributions are considered as the prior distributions for the entries in $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ for a given N: $\lambda_j \sim Gamma(\alpha_j, \beta_j)$, $j = 1, 2, \cdots, N-1$, and $\mu_j \sim Gamma(\gamma_j, \sigma_j)$, $j = 1, 2, \cdots, N$, where α_j, β_j, γ_j and σ_j are the hyperparameters. The prior distribution $p(N)$ of N can be specified based on the problem and available information. An appropriate candidate for the distribution of N is a non-informative discrete uniform distribution with a desired upper bound. To differentiate the parameter sets for different values of N, we denote $\boldsymbol{\Theta}_N = [\boldsymbol{\lambda}, \boldsymbol{\mu}]$. Then, the posterior distribution can be obtained as:

$$\begin{aligned} &p(\boldsymbol{\Theta}_N, N|(\boldsymbol{t}, \boldsymbol{n})) \propto \mathcal{L}(\boldsymbol{\Theta}_N, N|(\boldsymbol{t}, \boldsymbol{n}))p(\boldsymbol{\Theta}_N|N)p(N) \\ &= \prod_{k=1}^{m} \boldsymbol{\pi} \boldsymbol{P}^{\star}_{t_k} \mathbf{1} \prod_{j=1}^{N-1} \frac{1}{\Gamma(\alpha_k)} \frac{1}{\beta_j^{\alpha_j}} \lambda_j^{\alpha_j - 1} \exp\left(-\frac{\lambda_j}{\beta_j}\right) \prod_{j=1}^{N} \frac{1}{\Gamma(\gamma_j)} \frac{1}{\sigma_j^{\gamma_j}} \mu_j^{\gamma_j - 1} \exp\left(-\frac{\mu_j}{\sigma_j}\right) \times p(N), \end{aligned} \tag{4.25}$$

where $\mathcal{L}(\boldsymbol{\Theta}_N, N|(\boldsymbol{t}, \boldsymbol{n}))$ is the likelihood function same as equation (4.16).

4.5.1.2 Case with covariates For a case with covariates, we will use the same life-stress relationship as used in section 4.4.2. Moreover, the prior distributions of the parameters N, $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ remain similar to the case without covariates. Depending on the problem an appropriate prior distribution should be selected for the vector of regression coefficients $\boldsymbol{b}$. Since the entries of $\boldsymbol{b}$ could be negative or positive, Normal and Uniform distributions can be used. Then, the posterior

distribution can be obtained as:

$$
\begin{aligned}
&p(\boldsymbol{\Theta}_N, N, \boldsymbol{b}|(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n})) \propto \mathcal{L}(\boldsymbol{\Theta}_N, N|(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n}))p(\boldsymbol{\Theta}_N|N)p(N)p(\boldsymbol{b}) \\
&= \prod_{k=1}^{m} \boldsymbol{\pi} \boldsymbol{P}_{t_k}^{\star} \mathbf{1} \prod_{j=1}^{N-1} \frac{1}{\Gamma(\alpha_j)} \frac{1}{\beta_j^{\alpha_j}} \lambda_j^{\alpha_j - 1} \exp\left(-\frac{\lambda_j}{\beta_j}\right) \prod_{j=1}^{N} \frac{1}{\Gamma(\gamma_j)} \frac{1}{\sigma_j^{\gamma_j}} \mu_j^{\gamma_j - 1} \exp\left(-\frac{\mu_j}{\sigma_j}\right) \times p(N) \times p(\boldsymbol{b}),
\end{aligned}
\tag{4.26}
$$

where $\mathcal{L}(\boldsymbol{\Theta}_N, N|(\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n}))$ is the likelihood function same as equation (4.23).

4.5.2 Reversible Jump Markov Chain Monte Carlo

Clearly, for either case, changing the value of N affects the number of model parameters in $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$. Using the RJMCMC method, one of the following three types of parameter updates is randomly selected and performed in each iteration of the algorithm:

- fixed dimension parameter update,
- split a phase into two or combine two phases into one,
- birth of a new phase or death of an existing phase.

Let $\mathbf{y} = (\boldsymbol{t}, \boldsymbol{n})$ or $\mathbf{y} = (\boldsymbol{t}, \boldsymbol{X}, \boldsymbol{n})$ for the case without or with covariates. Specially, for the case without covariates, each move is accepted with probability $min(AcR, 1)$ where AcR is calculated by:

$$
AcR = \frac{\mathcal{L}(\Theta_{N^*}, N^*|\mathbf{y})p(\Theta_{N^*}|N^*)p(N^*)}{\mathcal{L}(\Theta_N, N|\mathbf{y})p(\Theta_N|N)p(N)} \times \frac{Q_{N^*,N}p(u^*, v^*|N^*, N, \Theta_{N^*})}{Q_{N,N^*}p(u, v|N, N^*, \Theta_N)} \times \left|\frac{\partial(\Theta_{N^*}, u^*, v^*)}{\partial(\Theta_N, u, v)}\right|,
\tag{4.27}
$$

in which N and N^* denote the current and the proposed number of phases; Q_{N,N^*} is the probability of moving from N phases to N^* phases, and $Q_{N^*,N}$ is that of moving from N^* phases to N phases; the last term is the Jacobian for the transformation of the parameters; u, v, u^*, and v^* are the auxiliary variables utilized to maintain the dimensionality of the existing and the proposed

parameter spaces, (Θ_N, u, v) and (Θ_{N^*}, u^*, v^*), respectively. To facilitate the search for the best value of N, a range of integer values for N, from 2 to a pre-specified N_{max}, can be considered. Similarly, for the case with covariates, equation (4.27) is simply modified by including the vector of regression coefficients $\boldsymbol{b}$ as part of model parameters.

In order to have a set of new parameters close enough to the existing parameters, one way is to establish a proposed set of parameters such that the mean time and probability of absorption do not change. Therefore, the following transformation is necessary for a phase that is making a split/combine or birth/death move:

$$\begin{aligned} \frac{\mu_j}{\mu_j+\lambda_j} &= \frac{\mu_a}{\mu_a+\lambda_a} + \left(\frac{\lambda_a}{\mu_a+\lambda_a} \times \frac{\mu_b}{\mu_b+\lambda_b}\right), \\ \frac{\mu_j}{\mu_j+\lambda_j} &= \frac{1}{\mu_a+\lambda_a} + \frac{1}{\mu_b+\lambda_b}, \end{aligned} \tag{4.28}$$

In particular, in split and birth moves, μ and λ indicate the current rates, and μ_a, μ_b, λ_a, and λ_b are the proposed rates; while in combine and death moves, μ_a, μ_b, λ_a, and λ_b indicate the current rates, and μ and λ show the proposed rates. For more information regarding the parameter updating process, readers are referred to the work of McGrory et al. (2009).

4.6 Numerical Examples

4.6.1 Simulation study

To illustrate the capability of the PH-based methods for analyzing censored aggregate failure-time data, data is generated from different distributions, including Gamma, Weibull, IG and Lognormal, and fitted with 3-phase Coxian distribution. From each distribution, three sets of data are generated with sizes of 12, 30 and 100 with a censoring time of 25500 hours. It is worth pointing out that in this simulation study, the individual failure times are made available so we can apply the Kaplan-Meier estimator to the individual failure times and compare the results with those obtained from the parametric models based on the censored aggregate data.

Figures 4.3 through 4.6 illustrate the estimates from the three-phase Coxian along with the

underlying distribution of the data as well as K-M estimator of individual data. In all four figures, as the number of data points increase, the estimated Coxian becomes closer and closer to the real underlying distribution. One can see that the Coxian-based method provides accurate results comparable to the Kaplan-Meier estimates on the individual failure times. Indeed, the PH-based method has the capability of altering the number of phases in favor of a desired performance and precision. This allows for its usage as a robust method to eliminate selecting other competitive parametric models when handling various underlying distributions.

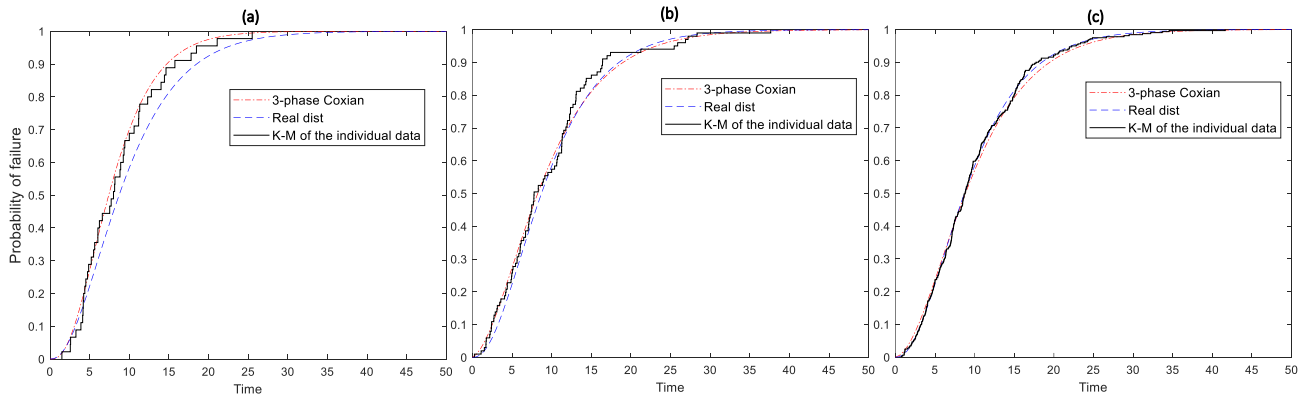

Figure 4.3: Estimated CDF of a 3-phase Coxian distribution from (a) 12, (b) 30, and (c) 100 simulated censored aggregate data points from $Gamma(2.5, 4)$ with a censoring time of 25500 h.

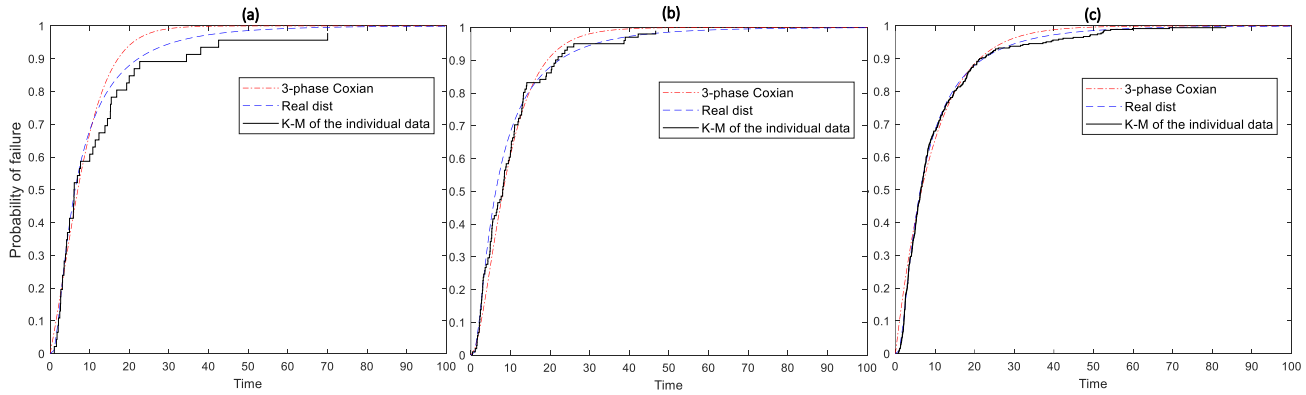

Figure 4.4: Estimated CDF of a 3-phase Coxian distribution based on (a) 12, (b) 30, and (c) 100 simulated censored aggregate data points from $IG(10, 8)$ with a censoring time of 25500 h.

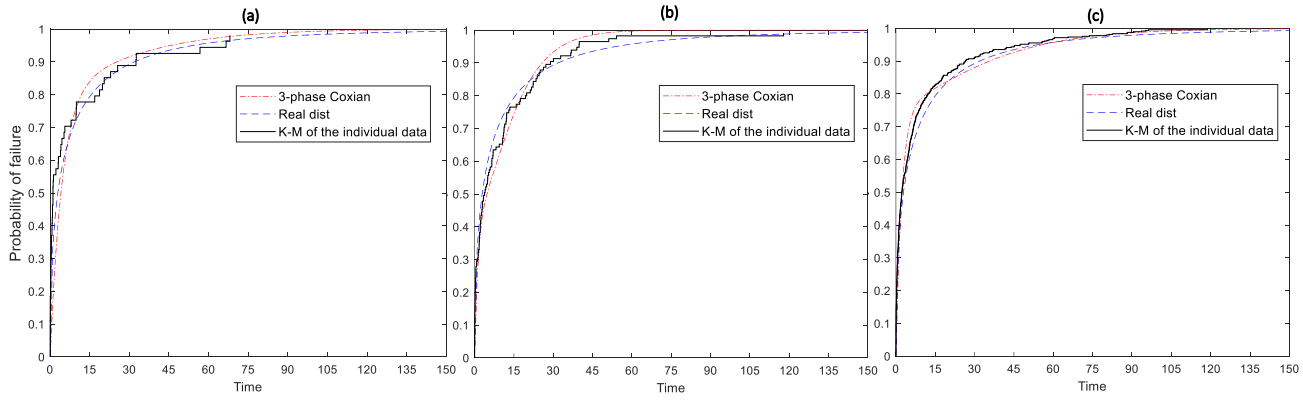

Figure 4.5: Estimated CDF of a 3-phase Coxian distribution based on (a) 12, (b) 30, and (c) 100 simulated censored aggregate data points from $Weibull(6, 0.5)$ with a censoring time of 25500 h.

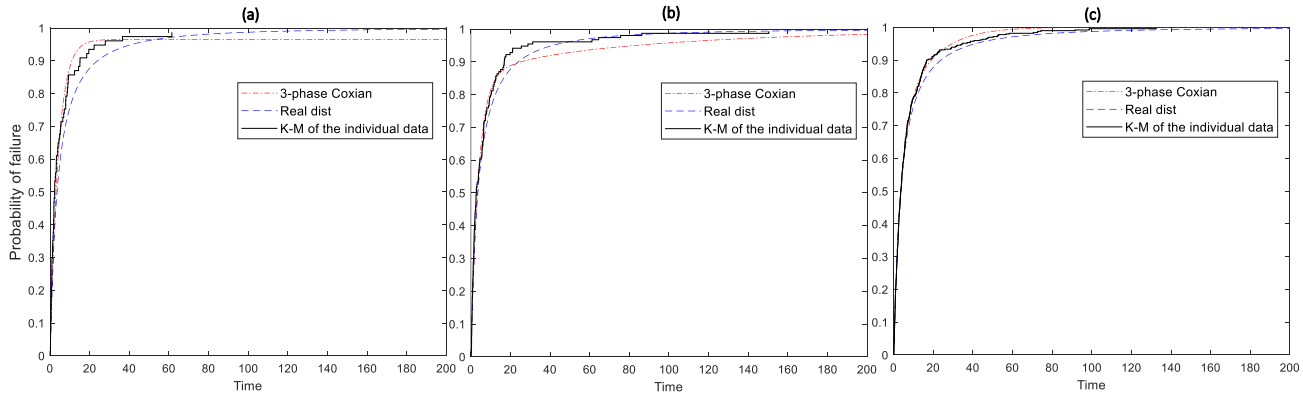

Figure 4.6: Estimated CDF of a 3-phase Coxian distribution from (a) 12, (b) 30, and (c) 100 simulated censored aggregate data points from $Lognormal(1.25, 1.5)$ with a censoring time of 25500 h.

4.6.2 Real data without covariates

The methods proposed in this work are tested on a set of censored aggregate data from Mahar et al. (2011). This dataset is supported by the Reliability Information Analysis Center (RIAC), which is a U.S. DoD's center of excellence in reliability, maintainability and quality.

The censored aggregate data presented in Table 4.1 is the recorded data from a certain component position of an electromagnetic relay in 12 army aircraft. The individual component failure times are unavailable. Instead, for each system (aircraft), the number of failed components or,

equivalently, the number of replacements in the time interval from the system installment to the inspection time (25500 hours) was recorded.

Table 4.1: Censored aggregate data of electromagnetic relays: n_k is the number of replacements in the k^{th} system and t_k is the length of time during which n_k replacements occurred.

n_k	3	1	3	1	1	9	2	1	1	5	1	1
t_k (1000 h)	25.5	25.5	25.5	25.5	25.5	25.5	25.5	25.5	25.5	25.5	25.5	25.5

In addition to the PH-based methods, two other distributions are also applied to the same dataset for comparison. Since the Gamma and IG distributions have closed-form expressions for convolutions, they can be used as feasible alternatives for the analysis of censored aggregate data. The likelihood functions of censored aggregate data for the Gamma and IG distributions are:

$$\begin{aligned} L_{Gamma}((\alpha,\theta)|(\boldsymbol{t},\boldsymbol{n})) &= \textstyle\prod_{k=1}^{m}(F(t_k|n(i)\alpha,\theta) - F(t_k|(n(i)+1)\alpha,\theta)) \\ &= \textstyle\prod_{k=1}^{m}(\frac{1}{\Gamma(n(i)\alpha)}\gamma(n(i)\alpha,\frac{x}{\theta}) - \frac{1}{\Gamma((n(i)+1)\alpha)}\gamma((n(i)+1)\alpha,\frac{x}{\theta})), \end{aligned} \tag{4.29}$$

and

$$\begin{aligned} L_{IG}((\mu,\lambda)|(\boldsymbol{t},\boldsymbol{n})) &= \textstyle\prod_{k=1}^{m}(F(t_k|n(i)\mu,n(i)^2\lambda) - F(t_k|(n(i)+1)\mu,(n(i)+1)^2\lambda)) \\ &= \textstyle\prod_{k=1}^{m}(\Phi(\sqrt{\frac{n(i)\lambda}{t(i)}}(\frac{t(i)}{n(i)^2\mu}-1)) + e^{\frac{2n(i)\lambda}{\mu}}\Phi(-\sqrt{\frac{n(i)\lambda}{t(i)}}(\frac{t(i)}{n(i)^2\mu})+1) \\ &-(\Phi(\sqrt{\frac{(n(i)+1)\lambda}{t(i)}}(\frac{t(i)}{(n(i)+1)^2\mu}-1)) + e^{\frac{2(n(i)+1)\lambda}{\mu}}\Phi(-\sqrt{\frac{(n(i)+1)\lambda}{t(i)}}(\frac{t(i)}{(n(i)+1)^2\mu})+1)), \end{aligned} \tag{4.30}$$

where α and θ are the shape and rate parameters of the Gamma distribution, and μ and λ are the parameters of of the IG distribution. The Gamma model was first introduced by Winkelmann (1995). The Gamma and IG models were used by Chen et al. (2020) under a Bayesian framework.

Next, the proposed PH models are applied. Specially, Coxian distributions with different numbers of phases are used, and a summary of the ML estimation results, including the parameter estimates and the corresponding likelihood values for the corresponding distributions are provided

in Table 4.2. Specially, for the 3-, 5-, and 7-phase Coxian distributions:

$$\hat{\mathbb{S}}_3 = \begin{pmatrix} -5.481 & 3.557 & 0 \\ 0 & -0.118 & 0.118 \\ 0 & 0 & -0.118 \end{pmatrix}, \quad \hat{\mathbb{S}}_5 = \begin{pmatrix} -7.021 & 4.380 & 0 & 0 & 0 \\ 0 & -0.240 & 0.240 & 0 & 0 \\ 0 & 0 & -0.240 & 0.240 & 0 \\ 0 & 0 & 0 & -0.240 & 0.240 \\ 0 & 0 & 0 & 0 & -0.240 \end{pmatrix},$$

$$\hat{\mathbb{S}}_7 = \begin{pmatrix} -8.856 & 5.469 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.362 & 0.362 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.362 & 0.362 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.362 & 0.362 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.362 & 0.362 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.362 & 0.362 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.362 \end{pmatrix}. \tag{4.31}$$

Table 4.2: The results of applying different models to the censored aggregate data.

Model	Gamma	IG	Coxian-3	Coxian-5	Coxian-7
Parameter estimates	$\hat{\alpha} = 0.4180$, $\hat{\theta} = 34.1463$	$\hat{\mu} = 3.4575e6$, $\hat{\lambda} = 1.8433$	$\hat{\mathbb{S}}_3$	$\hat{\mathbb{S}}_5$	$\hat{\mathbb{S}}_7$
Log-likelihood	-24.0551	-24.5080	-23.3064	-22.6621	-22.2306

Based on the likelihood values in Table 4.2, one can see that as the number of phases increases, the quality of the estimation improves due to the capability of PH distribution in mimicking the distributions of nonnegative random variables. An interesting view is that an N-phase Coxian distribution has $2N+1$ parameters in principle, but the EM algorithm results in a one phase of Coxian plus an $(N-1)$-phase Erlang. Indeed, the number of estimated parameters is 3 for the three Coxian alternatives. In other words, in this application, the PH-based method provides high estimation accuracy with the same number of parameters. In practice, the model with an intermediate number of phases is suggested, such as the 3- or 5-phase Coxian in this example. Model selection procedure can be tricky for Phase-type distribution. Because of the special structure of PH distribution, AIC penalty is too heavy. Phase-type distribution, in practice, has a flexible number of parameters. For instance, in a 5-phase Coxian distribution there are 11 parameters. But in the case of RIAC data, fitting a 5-phase Coxian results in a 3-parameter distribution which is much less than 11. There-

fore, the number of parameters varies case by case. AIC as well as the other alternative, BIC, are not appropriate options for comparisons of Phase-type distribution fitting. Although some cases where enough data is available, AIC and BIC may show the strength of the model. The issue of number of data points has been referenced the method "Sample size Adjusted Bayesian Information Criterion" (SABIC) (Sclove, 1987). This option may be useful in some cases, however, we need a minimum data size of 23 in order not to get a negative penalty on the number of parameters. In the RIAC data set, we have 12 data points. Figure 4.7 shows the CDF estimates using the ML estimation method to compare the three different parametric models.

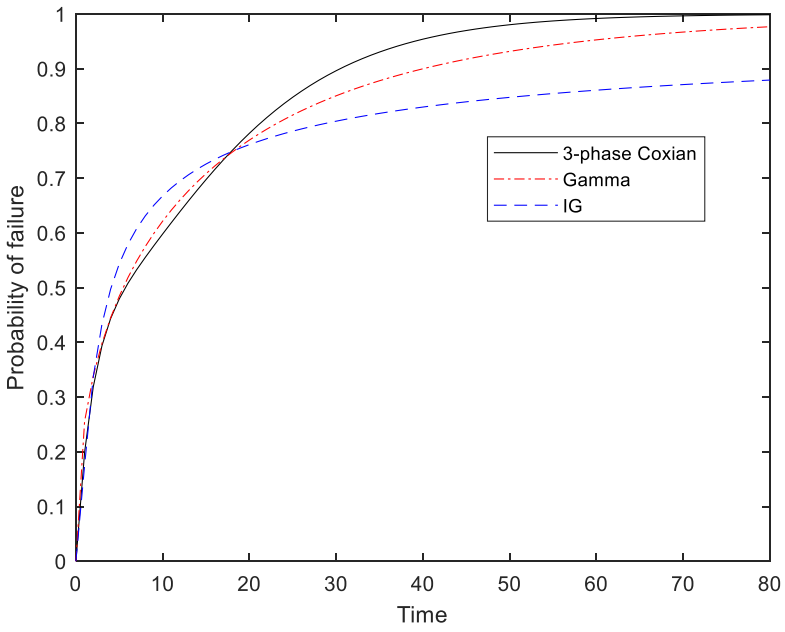

Figure 4.7: ML estimates of CDF by the 3-phase Coxian, Gamma and IG distributions.

The proposed PH-based Bayesian method is also performed on the same dataset. Note that the PH-based Bayesian method is capable of performing automatic model selection. Table 4.3 shows the results of 4-, 6-, and 8-phase Coxian distributions as well as those of Gamma and IG performed by Chen et al. (2020). The values for the three Coxian distributions are the results of three different runs of the algorithm started with an initial 3-phase Coxian model. In Figure 4.8, 90% point-wise credible intervals for the CDF of failure time are provided for the estimated Coxian

distributions. Since the extra parameters result in a higher Bayesian evidence, by considering the structural complexity as an additional criterion, the 6-phase Coxian distribution is used as the final model for the Bayesian method.

Table 4.3: The results of Bayesian evidence of the Gamma, IG and Coxian models for the censored aggregate data.

Model	Gamma	IG	Coxian-4	Coxian-6	Coxian-8
Log of Bayesian Evidence	-27.88	-33.90	-18.1021	-14.6134	-10.5914

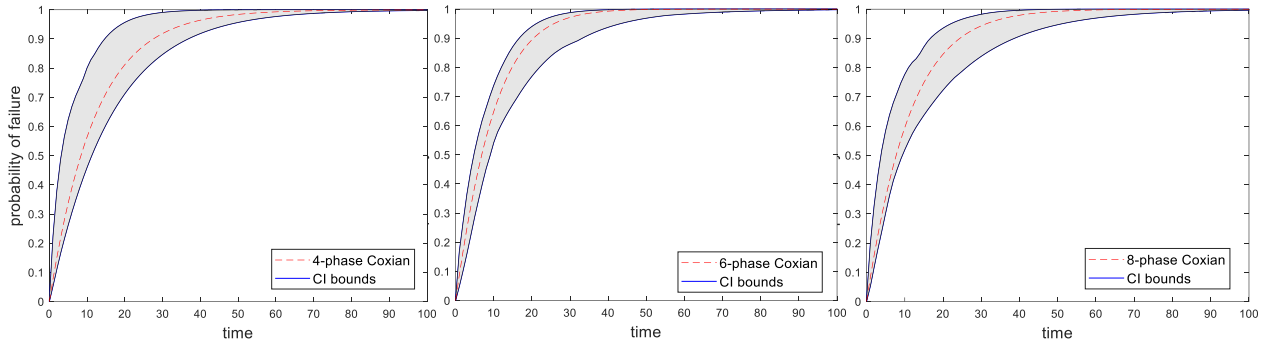

Figure 4.8: 90% credible intervals of CDF of failure time from three different runs of the Bayesian PH-based method for the censored aggregate data.

The model was tuned to produce an acceptance rate of $20\% - 25\%$. As in an RJMCMC model a number of different models are included, usual convergence criteria is not usable. Therefore, methods have been created specifically to monitor the convergence of RJMCMC models. One solution is to keep track of a parameter that uniquely defines the model and examine the parameter's convergence (Brooks et al., 2003). Notice that this method could underestimate the convergence time. Regarding the appropriate number of iterations, since there is no certain method introduced in the literature, we used trial and error to determine the sufficient number of iterations.

For the parameter uniquely identifying the model, we study the first three moments of the distribution as they can almost uniquely determine the shape of the distribution. Figure 4.9 illustrates the first moment (mean), the second central moment (variance), the third standardized moment (skewness) of the estimated Coxian distribution as well as the log-posterior in each iteration. The

log-posterior is calculated by substituting the point estimates of parameters in each iteration in the natural logarithm of equation (4.25). One can see that the convergence of these quantities is obvious after 1700 iterations for the RIAC data set.

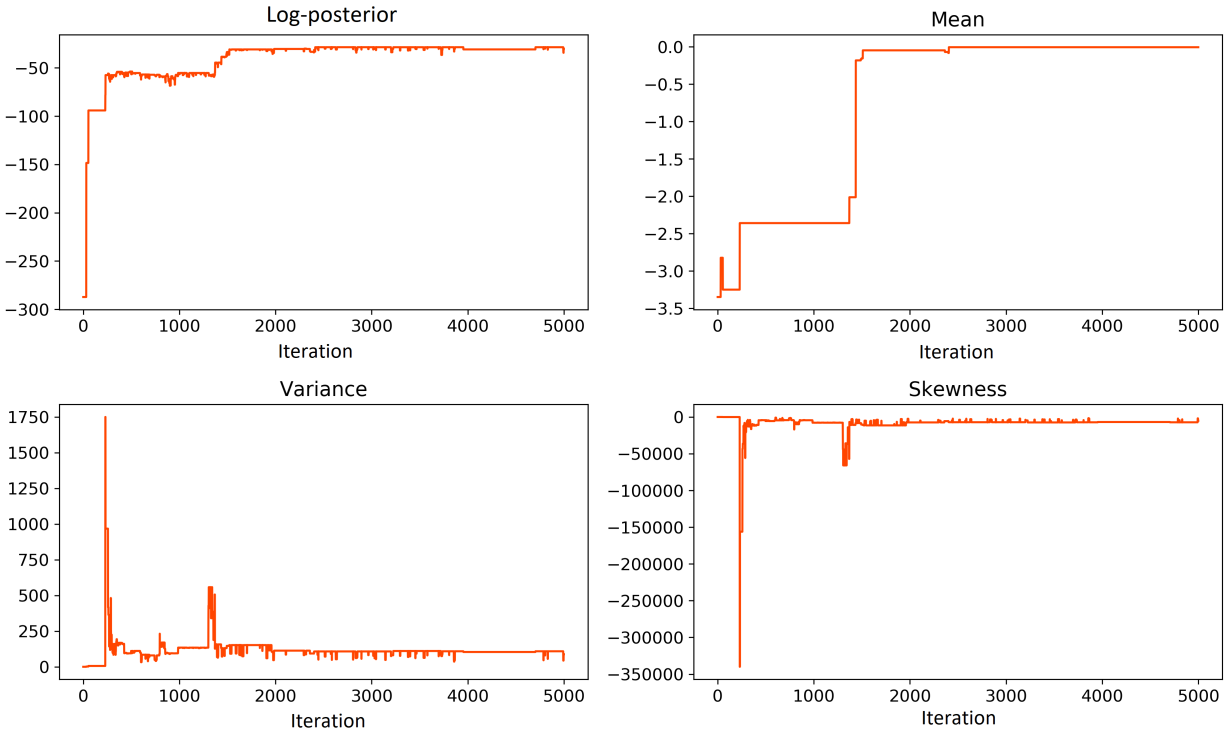

Figure 4.9: Convergence of log-posterior, and the mean, variance and skewness of the estimated Coxian distribution.

4.6.3 Real data with covariates

To test the strength of the proposed methods in handling censored aggregate data with covariates, a set of accelerated life testing (ALT) data from a type of miniature lamps with voltage as the covariate is used (Liao & Karimi, 2017). This dataset contains individual failure times of 77 components as well as censoring time for 17 components under three different voltage levels. The original dataset is presented in Table 4.4.

Table 4.4: ALT data of miniature lamps.

Stress	Lifetime in hours ("+" censored)							
5V ($z_1 = 1$)	20.5	22.3	23.2	24.7	26	34.1	39.6	41.8
	43.6	44.9	47.7	61.6	62.1	65.5	70.8	87.8
	118.3	120.1	145.4	157.4	180.9	187.7	204	206.7
	213.9	215.2	218.7	254.1	262.6	293	304	313.7
	314.1	317.9	337.7	430.2				
3.5V ($z_2 = 0.5$)	37.8	43.6	51.1	58.6	65.5	65.9	75.6	82.5
	88.1	89	106.6	113.1	121.1	121.5	128.3	151.8
	171.7	181	202.7	211.7	230.7	249.9	275.6	285
	296.2	358.5	379.8	434.5	493.1	506.1	570	577.7
	876.3	890+	890+	922	941+	941+		
2V ($z_3 = 0$)	223.1	254	316.7	560.2	679	737	894.4	930.5+
	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+
	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+	930.5+

To illustrate how the proposed PH-based models perform, the data is combined into a set of censored aggregate data as shown in Table 4.5. The advantage of using this dataset is that the individual failure times can be used for comparison. Specifically, our estimated CDF can be compared with the Kaplan-Meier estimate at each stress level. In addition to the Coxian distributions with different numbers of phases, for comparison, the Gamma and IG distributions are also considered with $\theta = \frac{1}{\alpha \exp(\boldsymbol{bX})}$ for Gamma and $\mu = \exp(\boldsymbol{bX})$ for IG. Note that only the ML method is applied in this example, but the Bayesian alternative can be applied too.

Table 4.5: Censored aggregated data from ALT data of miniature lamps.

Stress		Lifetime in hours ("+" censored)						
5V ($z_1 = 1$)	time	450	450	450	450	450	450	450
	no. fails	2	3	4	2	4	1	3
	time	450	450	450				
	no. fails	2	2	3				
3.5V ($z_2 = 0.5$)	time	1000	1000	1000	1000	1000	1000	1000
	no. fails	4	3	2	3	3	7	2
	time	500	890	890	941	941		
	no. fails	2	0	0	0	0		
2V ($z_3 = 0$)	time	1500	1000	700	930.5	930.5	930.5	930.5
	no. fails	1	2	1	0	0	0	0
	time	930.5	930.5	930.5	930.5	930.5	930.5	930.5
	no. fails	0	0	0	0	0	0	0
	time	930.5	930.5	930.5	930.5	930.5	930.5	
	no. fails	0	0	0	0	0	0	

The ML estimation results, including the parameter estimates and the corresponding likelihood values for the corresponding distributions are provided in Table 4.6. Specially, for the 3-, 5-, and 7-phase Coxian distributions:

$$\hat{\mathbb{S}}_3 = \begin{pmatrix} -0.00380 & 0.00160 & 0 \\ 0 & -0.00380 & 0.00337 \\ 0 & 0 & -0.00380 \end{pmatrix}, \quad \hat{\mathbb{S}}_5 = \begin{pmatrix} -0.00916 & 0.00497 & 0 & 0 & 0 \\ 0 & -0.00916 & 0.00756 & 0 & 0 \\ 0 & 0 & -0.00916 & 0.00811 & 0 \\ 0 & 0 & 0 & -0.00916 & 0.00785 \\ 0 & 0 & 0 & 0 & -0.00916 \end{pmatrix},$$

$$\hat{\mathbb{S}}_7 = \begin{pmatrix} -0.01534 & 0.00944 & 0 & 0 & 0 & 0 & 0 \\ 0 & -0.01534 & 0.01091 & 0 & 0 & 0 & 0 \\ 0 & 0 & -0.01534 & 0.01314 & 0 & 0 & 0 \\ 0 & 0 & 0 & -0.01534 & 0.01510 & 0 & 0 \\ 0 & 0 & 0 & 0 & -0.01534 & 0.01526 & 0 \\ 0 & 0 & 0 & 0 & 0 & -0.01534 & 0.01497 \\ 0 & 0 & 0 & 0 & 0 & 0 & -0.01534 \end{pmatrix}.$$

One can see that the Coxian distributions outperform the Gamma and IG distributions in terms of the likelihood values. Moreover, the ML method produces Coxian structures as much as needed with the rest of the CTMC being Erlang. By increasing the number of phases in the Coxian

distribution, the likelihood value increases. To balance the complexity of the model and estimation accuracy, the 3-phase Coxian model is selected in this application. Figure 4.10 shows the CDF estimates for the three stress levels using the Kaplan-Meier estimator based on the original ALT data and the three parametric models based the censored aggregate data. Note that the estimates from the 3-phase Coxian model match the Kaplan-Meier estimates for the first and third stress levels, but for the second stress level, the result deviates from the Kaplan-Meier estimate. The problem lies in the significant information loss when aggregating the individual ALT data into the censored aggregate data.

Table 4.6: The results of different models on the censored aggregate data given in Table 4.5.

Model	Gamma	IG	Coxian-3	Coxian-5	Coxian-7
Parameter estimates	$\hat{\alpha} = 0.886$, $\hat{b_1} = -7.902$, $\hat{b_2} = 2.870$	$\hat{\lambda} = 376.8359$, $\hat{b_1} = -8.195$, $\hat{b_2} = 2.784$	$\hat{\mathbb{S}}_3$, $\hat{b_1} = -1.600$, $\hat{b_2} = 2.775$	$\hat{\mathbb{S}}_5$, $\hat{b_1} = -2.152$, $\hat{b_2} = 2.882$	$\hat{\mathbb{S}}_7$, $\hat{b_1} = -2.581$, $\hat{b_2} = 3.083$
Log-likelihood	-53.823	-58.854	-53.6954	-53.4583	-53.2237

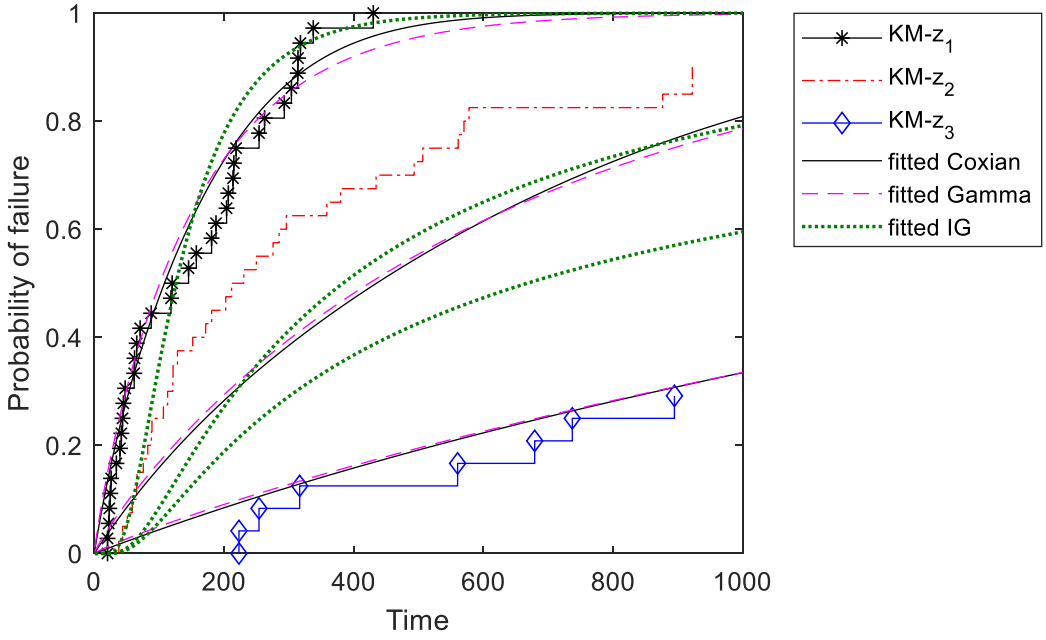

Figure 4.10: Comparison of statistical fits of the three parametric models capable of handling censor aggregate data.

4.7 Conclusions

This work uses a PH distribution model to analyze censored aggregate failure-time data for the first time. To estimate the model parameters, an EM algorithm is developed for the ML estimation method and an alternative Bayesian method is also provided. The simulation study shows the flexibility of PH distribution in approximating other probability distributions widely used in reliability engineering. Moreover, the two real-world examples demonstrate the capability of the PH-based method in the analysis of censored aggregate failure-time data with and without covariates. It is worth pointing out that the proposed method provides a new direction for analyzing such data generated from a variety of nonnegative random variables. From a technical point of view, this method can be used as a powerful substitute for those underlying probability distributions that are mathematically intractable in analyzing such data. For a practitioner, it alleviates the burden of model

selection and provides great flexibility in determining the model complexity and interpretability based on computational challenges and other needs.

References

Asmussen, S., Jensen, J. L., & Rojas-Nandayapa, L. (2016). On the Laplace transform of the lognormal distribution. *Methodology and Computing in Applied Probability*, *18*(2), 441–458.

Asmussen, S., Nerman, O., & Olsson, M. (1996). Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics*, *23*(4), 419–441.

Ausıén, M. C., Wiper, M. P., & Lillo, R. E. (2008). Bayesian prediction of the transient behaviour and busy period in short-and long-tailed GI/G/1 queueing systems. *Computational Statistics & Data Analysis*, *52*(3), 1615–1635.

Beaulieu, N. C., & Rajwani, F. (2004). Highly accurate simple closed-form approximations to lognormal sum distributions and densities. *IEEE Communications Letters*, *8*(12), 709–711.

Beaulieu, N. C., & Xie, Q. (2004). An optimal lognormal approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology*, *53*(2), 479–489.

Brooks, S., Giudici, P., & Philippe, A. (2003). Nonparametric convergence assessment for MCMC model selection. *Journal of Computational and Graphical Statistics*, *12*(1), 1–22.

Bučar, T., Nagode, M., & Fajdiga, M. (2004). Reliability approximation using finite Weibull mixture distributions. *Reliability Engineering & System Safety*, *84*(3), 241–251.

Buchholz, P., Kriege, J., & Felko, I. (2014). *Input Modeling with Phase-type Distributions and Markov Models: Theory and Applications*. Springer, New York.

Cameron, A. C., & Johansson, P. (1997). Count data regression using series expansions: with applications. *Journal of Applied Econometrics*, *12*(3), 203–223.

Chen, C.-M., Chuang, Y.-W., & Shen, P.-S. (2015). Two-stage estimation for multivariate recurrent event data with a dependent terminal event. *Biometrical Journal*, *57*(2), 215–233.

Chen, P., & Ye, Z.-S. (2017). Estimation of field reliability based on aggregate lifetime data. *Technometrics*, *59*(1), 115–125.

Chen, P., Ye, Z.-S., & Zhai, Q. (2020). Parametric analysis of time-censored aggregate lifetime data. *IISE Transactions*, *52*(5), 516–527.

Coit, D. W., & Dey, K. A. (1999). Analysis of grouped data from field-failure reporting systems. *Reliability Engineering & System Safety*, *65*(2), 95–101.

Coit, D. W., & Jin, T. (2000). Gamma distribution parameter estimation for field reliability data with missing failure times. *IIE Transactions*, *32*(12), 1161–1166.

Denson, W., Crowell, W., Jaworski, P., & Mahar, D. (2014). *Electronic Parts Reliability Data 2014*. Reliability Information Analysis Center, Rome, NY.

Elmahdy, E. E., & Aboutahoun, A. W. (2013). A new approach for parameter estimation of finite Weibull mixture distributions for reliability modeling. *Applied Mathematical Modelling*, *37*(4), 1800–1810.

Gu, W., Fan, N., & Liao, H. (2019). Evaluating readmission rates and discharge planning by analyzing the length-of-stay of patients. *Annals of Operations Research*, *276*(1–2), 89–108.

He, S., Zhang, Z., Jiang, W., & Bian, D. (2018). Predicting field reliability based on two-dimensional warranty data with learning effects. *Journal of Quality Technology*, *50*(2), 198–206.

Horváth, A., & Telek, M. (2007). Matching more than three moments with acyclic phase type distributions. *Stochastic Models*, *23*(2), 167–194.

Jin, T., & Gonigunta, L. S. (2010). Exponential approximation to Weibull renewal with decreasing failure rate. *Journal of Statistical Computation and Simulation*, *80*(3), 273–285.

Karimi, S., Liao, H., & Fan, N. (2020). Flexible methods for reliability estimation using aggregate failure-time data. *IISE Transactions*, (in print).

Kharrat, T., Boshnakov, G. N., McHale, I., & Baker, R. (2019). Flexible regression models for count data based on renewal processes: The Countr package. *Journal of Statistical Software*, *90*(13), 1–35.

Lam, C. L. J., & Le-Ngoc, T. (2007). Log-shifted gamma approximation to lognormal sum distributions. *IEEE Transactions on Vehicular Technology*, *56*(4), 2121–2129.

Liao, H., & Karimi, S. Comparison study on general methods for modeling lifetime data with covariates. In: *2017 Prognostics and System Health Management Conference (PHM-Harbin)*. IEEE. Harbin, China, 2017, 1–5.

Liu, X., & Pan, R. (2020). Analysis of Large Heterogeneous Repairable System Reliability Data with Static System Attributes and Dynamic Sensor Measurement in Big Data Environment. *Technometrics*, *62*(2), 206–222.

Mahar, D., Fields, W., Reade, J., Zarubin, P., & McCombie, S. (2011). *Nonelectronic parts reliability data*. Reliability Information Analysis Center, Rome, NY.

McGrory, C. A., Pettitt, A. N., & Faddy, M. J. (2009). A fully Bayesian approach to inference for Coxian phase-type distributions with covariate dependent mean. *Computational Statistics & Data Analysis*, *53*(12), 4311–4321.

McShane, B., Adrian, M., Bradlow, E. T., & Fader, P. S. (2008). Count models based on Weibull interarrival times. *Journal of Business & Economic Statistics*, *26*(3), 369–378.

Meeker, W. Q., & Hong, Y. (2014). Reliability meets big data: opportunities and challenges. *Quality Engineering*, *26*(1), 102–116.

Mehta, N. B., Wu, J., Molisch, A. F., & Zhang, J. (2007). Approximating a sum of random variables with a lognormal. *IEEE Transactions on Wireless Communications*, *6*(7), 2690–2699.

Okamura, H., Watanabe, R., & Dohi, T. (2014). Variational Bayes for phase-type distribution. *Communications in Statistics-Simulation and Computation*, *43*(8), 2031–2044.

OREDA. (2009). *OREDA Offshore Reliability Data Handbook*. Det Norske Veritas, Norway.

Osogami, T., & Harchol-Balter, M. (2003). A closed-form solution for mapping general distributions to minimal PH distributions. *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, 200–217.

Osogami, T., & Harchol-Balter, M. (2006). Closed form solutions for mapping general distributions to quasi-minimal PH distributions. *Performance Evaluation*, *63*(6), 524–552.

Riska, A., Diev, V., & Smirni, E. (2004). An EM-based technique for approximating long-tailed data sets with PH distributions. *Performance Evaluation*, *55*(1-2), 147–164.

Rizopoulos, D., Verbeke, G., & Lesaffre, E. (2009). Fully exponential Laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *71*(3), 637–654.

Rue, H., Martino, S., & Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series b (Statistical Methodology)*, *71*(2), 319–392.

Sclove, S. L. (1987). Application of Model-Selection Criteria to Some Problems in Multivariate Analysis. *Psychometrika*, *52*(03), 333–343.

Starobinski, D., & Sidi, M. (2000). Modeling and analysis of power-tail distributions via classical teletraffic methods. *Queueing Systems*, *36*(1-3), 243–267.

Varmazyar, M., Akhavan-Tabatabaei, R., Salmasi, N., & Modarres, M. (2020). Operating room scheduling problem under uncertainty: Application of continuous phase-type distributions. *IISE Transactions*, *52*(2), 216–235.

Watanabe, R., Okamura, H., & Dohi, T. (2012). An efficient MCMC algorithm for continuous PH distributions. *Proceedings of the 2012 Winter Simulation Conference (WSC)*, 1–12.

Winkelmann, R. (1995). Duration dependence and dispersion in count-data models. *Journal of Business & Economic Statistics*, *13*(4), 467–474.

Xiao, S., Kottas, A., Sansó, B., & Kim, H. (2020). Nonparametric Bayesian Modeling and Estimation for Renewal Processes. *Technometrics*, *62*(1), 1–16.

Yamaguchi, Y., Okamura, H., & Dohi, T. (2010). A variational Bayesian approach for estimating parameters of a mixture of Erlang distribution. *Communications in Statistics-Theory and Methods*, *39*(13), 2333–2350.

Yin, J., Zhu, X., & Huang, Y. (2016). 3D Markov chain based narrowband interference model for in-home broadband power line communication. *2016 IEEE Global Communications Conference (GLOBECOM)*, 1–6.

Chapter 4 Appendix

The conditional expectations of hidden variables of PH distribution, B_i, Z_i, N_{ij}, $i, j = 1, \cdots, N$ are derived for a single censored aggregate data point. Note that in the following equations, i refers to the ith phase of each individual component and v determines the component that we are referring to. For example, if $N = 3$ for a data point with $n = 4$, the corresponding PH process has $3(4 + 1) = 15$ phases. Given $i = 2$, we are referring to the phases 2, 5, 8, 11, 14 of the PH process or equivalently $i + N(v - 1)$. $\boldsymbol{\pi}$ and $\boldsymbol{P}$ are the initial probability vector and the transition matrix corresponding to data point (t, n). J_u indicates the state of the process at time u. To simplify the notation, $P_r(T \in dt)$ is used instead of $P_r(T \in dt|(\boldsymbol{\pi}, \mathbb{S}))$. The expected sojourn time in phase i is defined as:

$$
\begin{aligned}
E[Z_i|T = t] &= E\big[\int_0^\infty 1_{J_u=i} du|T = t\big] = \int_0^\infty P_r(J_u = i|T = t)du \\
&= \int_0^\infty \frac{P_r(J_u = i, T \in dt)}{P_r(T \in dt)} du = \frac{\int_0^\infty P_r(J_u = i)P_r(T \in dt|J_u = i)du}{P_r(T \in dt)} \\
&= \frac{1}{n+1} \frac{\int_0^t \sum_{v=1}^{n+1} \boldsymbol{\pi} e^{\mathbb{S}u} \boldsymbol{e}_{i+N(v-1)} \boldsymbol{e}'_{i+N(v-1)} e^{\mathbb{S}(t-u)} \boldsymbol{e}_{nN+1:(n+1)N} du}{\boldsymbol{\pi} e^{\mathbb{S}t} \boldsymbol{e}_{nN+1:(n+1)N}}, \quad i = 1, 2, \cdots, N.
\end{aligned}
$$

For the number of jumps between non-absorbing phases i and j, we first define $\epsilon > 0$. $N_{ij}^\epsilon = \sum_{k=0}^\infty \mathbf{1}_{J_{k\epsilon}=i, J_{(k+1)\epsilon=j}}$ gives a discrete approximation of the number of jumps (Asmussen et al. (1996)). As $\epsilon \to 0$, integration can be used to substitute the summation. We also note that:

$$
\frac{e^{\mathbb{S}\epsilon} - \boldsymbol{I}}{\epsilon} \to \mathbb{S} \text{ as } \epsilon \to 0.
$$

As a result, we can define the expected number of jumps as:

$$
\begin{aligned}
E[N_{ij}^{\epsilon}|T=t] &= \sum_{k=0}^{[t/\epsilon]-1} \frac{P_r(J_{k\epsilon}=i, J_{(k+1)\epsilon}=j, T\in dt)}{P_r(T\in dt)} \\
&= \sum_{k=0}^{[t/\epsilon]-1} \frac{P_r(J_{k\epsilon}=i)P_r(J_{(k+1)\epsilon}=j|J_{k\epsilon}=i)P_r(T\in dt|J_{(k+1)\epsilon}=j)}{P_r(T\in dt)} \\
&= \frac{1}{n+1}\frac{\sum_{k=0}^{[t/\epsilon]-1}\sum_{v=1}^{n+1}(\boldsymbol{\pi}e^{\mathbb{S}k\epsilon}\boldsymbol{e}_{i+(N(v-1))})(\mathbb{S}_{ij})(\boldsymbol{e}'_{j+N(v-1)}e^{\mathbb{S}(t-(k+1)\epsilon)}\boldsymbol{e}_{nN+1:(n+1)N})}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}} \\
&\to \frac{1}{n+1}\frac{\int_0^t\sum_{v=1}^{n+1}(\boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{i+(N(v-1))})(\mathbb{S}_{ij})(\boldsymbol{e}'_{j+N(v-1)}e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N})du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}.
\end{aligned}
$$

As can be conjectured from the structure of the censored aggregate data, we also have:.

$$
\begin{aligned}
E[N_{iN+1}^{\epsilon}|T=t] &= \sum_{k=0}^{[t/\epsilon]-1} \frac{P_r(J_{k\epsilon}=i, J_{(k+1)\epsilon}=abs, T\in dt)}{P_r(T\in dt)} \\
&= \sum_{k=0}^{[t/\epsilon]-1} \frac{P_r(J_{k\epsilon}=i)P_r(J_{(k+1)\epsilon}=abs|J_{k\epsilon}=i)P_r(T\in dt|J_{(k+1)\epsilon}=abs)}{P_r(T\in dt)} \\
&= \frac{1}{n+1}\frac{\sum_{k=0}^{[t/\epsilon]-1}\sum_{v=1}^{n}(\boldsymbol{\pi}e^{\mathbb{S}k\epsilon}\boldsymbol{e}_{i+(N(v-1))})(\mathbf{S}_i)(\boldsymbol{\pi}_{v+1}e^{\mathbb{S}(t-(k+1)\epsilon)}\boldsymbol{e}_{nN+1:(n+1)N})}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}} \\
&\to \frac{1}{n+1}\frac{\int_0^t\sum_{v=1}^{n}(\boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{i+(N(v-1))})(\mathbf{S}_i)(\boldsymbol{\pi}_{v+1}e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N})du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}},
\end{aligned}
$$

where abs means absorption. However, for this type of data, absorption causes immediate start of the life of next component and hence moving to one of the N phases related to the new component. Also, it is required to define an initial probability vector for each individual component. For component v, $\boldsymbol{\pi}_v$ will be a $1\times N(n+1)$ vector with all zero entries except the N positions from $(v-1)N+1$ to vN that are equal to the individual initial probability vector.

The expected number of times the process starts in the ith phase is more complicated. Indeed, for the first component, we know that the process starts at time 0, but the rest of components start their operations based on a probability function in the time interval $[0,t]$. Specially, for the first

component we have:

$$
\begin{aligned}
E_1[B_i|T=t] &= \frac{P_r(J_0=i, T\in dt)}{P_r(T\in dt)} \\
&= \frac{P_r(J_0=i)P_r(T\in dt|J_0=i)}{P_r(T\in dt)} \\
&= \frac{\boldsymbol{\pi}(i)\boldsymbol{e}_i^{'}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}.
\end{aligned}
$$

For the rest of components, we have:

$$
\begin{aligned}
E_{2:n+1}[B_i|T=t] &= \frac{\sum_{k=0}^{[t/\epsilon]-1}\sum_{v=1}^{n} P_r(J_{k\epsilon}=(v-1)N+1:vN, J_{(k+1)\epsilon=i+vN}, T\in dt)}{P_r(T\in dt)} \\
&= \sum_{k=0}^{[t/\epsilon]-1}\sum_{v=1}^{n}\Big(\frac{P_r(J_{k\epsilon}=(v-1)N+1:vN)}{P_r(T\in dt)} \\
&\times \frac{P_r(J_{(k+1)\epsilon}=i+vN|J_{k\epsilon}=(v-1)N+1:vN)P_r(T\in dt|J_{(k+1)\epsilon}=i+vN)}{P_r(T\in dt)}\Big) \\
&= \frac{\sum_{k=0}^{[t/\epsilon]-1}\sum_{v=1}^{n}\boldsymbol{\pi}e^{\mathbb{S}k\epsilon}\boldsymbol{e}_{(v-1)N+1:vN}\boldsymbol{S}_{v-1}\boldsymbol{\pi}(i)e^{\mathbb{S}(t-(k+1)\epsilon)}\boldsymbol{e}_{nN+1:(n+1)N}}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}} \\
&\to \frac{\int_0^t\sum_{v=1}^{n}\boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{(v-1)N+1:vN}\boldsymbol{S}_{v-1}\boldsymbol{\pi}(i)e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N}du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}.
\end{aligned}
$$

Therefore:

$$
\begin{aligned}
\Longrightarrow E[B_i|T=t] = \frac{1}{n+1}\Big(&\frac{\boldsymbol{\pi}(i)\boldsymbol{e}_i^{'}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}} \\
&+ \frac{\int_0^t\sum_{v=1}^{n}\boldsymbol{\pi}e^{\mathbb{S}u}\boldsymbol{e}_{(v-1)N+1:vN}\boldsymbol{S}_{v-1}\boldsymbol{\pi}(i)e^{\mathbb{S}(t-u)}\boldsymbol{e}_{nN+1:(n+1)N}du}{\boldsymbol{\pi}e^{\mathbb{S}t}\boldsymbol{e}_{nN+1:(n+1)N}}\Big).
\end{aligned}
$$

5 Data Selection from Large Data Sets for Limited Computational Resources

The challenge of dealing with large-scale data has been problematic in many data analysis applications such as healthcare. To handle such data, it is convenient to choose a subset of the data. However, it is extremely important how to select the subset. We propose a data selection method based on maximizing the information gain. The original optimization problem is a mixed integer linear programming (MINLP) with a highly nonlinear objective function. A method based on Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) clustering is proposed to mitigate the computational burden.

5.1 Introduction

5.1.1 Background and Motivation

Data selection has been investigated in several areas of application. On the one hand training data selection for machine learning methods including regression (Shanks, 2017), Natural Language Processing (NLP) (van der Wees et al., 2017), Support Vector Machines (SVM) (Kawulok & Nalepa, 2012) has been an area of research. On the other hand a wide variety of application-specific methods have been proposed including software cross-projec defect prediction (Herbold, 2013), energy consumption prediction (Paudel et al., 2017), potential energy surface of chemical systems (Guan et al., 2018), Alzheimer's disease (Khan et al., 2019), to name a few.

The data selection methods in the literature are generally very dependent on the application. However, a few works are more general. For example, Kawulok and Nalepa, 2012 uses a genetic algorithm for selecting data from large, noisy data sets for SVM, a statistical method with many applications. While there has been much more research on data selection for machine translation purposes, due to the frequent necessity, this problem has rarely been investigated in other applications, specifically reliability. In many applications of data analytics including reliability and healthcare, phase-type distributions are of great importance. To cope with the problem of big data for limited computational resources, in this chapter we investigate the case of Erlang distributed

data, which is a special type of phase-type distribution.

5.1.2 Overview

The rest of this work is organized as follows. Section 5.2 explains the preliminaries of this research work. Section 5.3 is dedicated to the proposed method. In section 5.4 a numerical example is provided to evaluate the method and its application in the real world. Lastly, conclusion and future research is provided in section 5.5.

5.2 Preliminaries on Erlang Distribution and Fisher Information

Erlang distribution is a special case of Phase-type distribution in which the jumps are only allowed to the immediate next phase and the sojourn rates are all similar, λ. The number of phases is shown by k. Figure 5.1 is a demonstration of Erlang distribution procedure. The pdf of Erlang distribution is as follows:

$$f(y) = \frac{\lambda^k y^{k-1} e^{-\lambda y}}{(k-1)!} \tag{5.1}$$

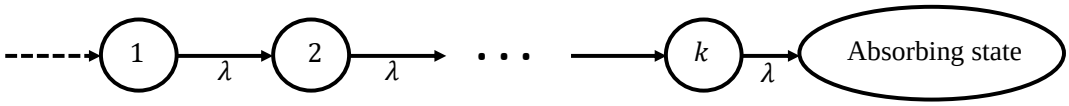

Figure 5.1: CTMC for k-Phase Erlang distribution.

To approximate information gain, determinant of Fisher information matrix is used. Fisher information matrix can be attained by:

$$FI = -\frac{\partial^2 L}{\partial \boldsymbol{\gamma}^T \partial \boldsymbol{\gamma}} \tag{5.2}$$

where L is the log-likelihood function and $\boldsymbol{\gamma}$ is the vector of distribution parameters.

Log-linear regression relation is used to explain the changes of the distribution parameter based on the covariates.

$$\lambda_i = exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_m x_{im}) \tag{5.3}$$

The parameter k is integer and therefore the derivative does not exist for it. So, we consider it a hyperparameter and will later apply cross-validation to get the estimate of k.

The following are the elements of Fisher information matrix for an Erlang distribution with log-linear regression relation:

$$\frac{\partial^2 L}{\partial \beta_j^2} = -\sum_{i=1}^{n} x_{ij}^2 y_i \lambda_i \tag{5.4}$$

$$\frac{\partial^2 L}{\partial \beta_j \partial \beta_l} = -\sum_{i=1}^{n} x_{ij} x_{il} y_i \lambda_i \tag{5.5}$$

Note that $x_{i0} = 1$ to be multiplied by the intercept.

5.3 Model

The goal is to select a subset of the data, based on calculation limitations, that maximizes the information. The calculation limitations could include the number of data points, the size of the subset of the data, etc. In this work, we consider a maximum allowed number of data points. To maximize information, the determinant of Fisher information matrix of the subset is used as the criterion.

Specially, the following objective function is considered:

$$\max_{\{C_j, K_j\}_{j=1}^{J}} |I_{\{C_j, K_j\}_{j=1}^{J}}|, C_j \in \{0, 1\}, K_j \in \{0, 1, 2, \cdots\} \tag{5.6}$$

where $|I_{\{C_j, K_j\}_{j=1}^{J}}|$ is the determinant of information matrix obtained based on the selected subset of the data. C_j shows level j, and K_j shows the number of data points selected from level j. It is worth pointing out that the objective function is highly nonlinear, which cannot be solved with

usual MINLP algorithms. To overcome this challenge, in this work, we have developed algorithms to deal with this problem and find the optimal solution efficiently.

Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) is a clustering algorithm that is well suited for handling large data sets (Zhang et al., 1997). BIRCH algorithm works based on CF-tree in-memory data structure. It can produce the results in only a few scans of the data due to it's hierarchical procedures. In this work, we apply Birch algorithm to the original data, producing c clusters. If the target number of choice of data points is n, we will take c/n data points from each cluster, resulting in a balanced subset of the data. The number of clusters c is another hyperparameter to be tuned. The c/n data points can be selected either randomly from each cluster or from the center of the cluster. The results from these choices will be studied and compared in a numerical example. It will be shown that the cluster center-based method improves the information gain significantly in comparison to random selection.

5.4 Numerical Study

To evaluate the effectiveness of the method, we applied the method to a set of data from *elecdemand* dataset from r *fpp2* package (Hyndman, 2020). The data represents the half-hourly operational electricity demand of Vicoria, Australia during 2014. The covariates are temperature and a categorical variable, workday, which is 0 on weekend days and 1 otherwise.

There are 17520 data points (days) and for each two covariates are recorded. First, an initial estimate of the model parameters based on an Erlang distribution with log-linear regression relation is acquired. Then, a maximum number of 3000 data points is determined as the choice of maximum allowed number of data.

The temperature varies between 1.5 to 43.2. To be able to assign the temperature covariate to levels, the following transformation was performed, where t is the temperature in a certain data point. Doing this, we will assign a range of 6 degrees Celsius to one value, getting a total of 6

levels for temperature.

$$t_{level} = \left[\frac{t - 1.5}{7}\right]. \tag{5.7}$$

Since the other covariate is work day, which is a binary variable, we will have a total of 12 levels.

In the next step, we need an initial guess about the parameters of Erlang distribution to find the the determinant of the Fisher Information. Based on the model assumption a maximum number of usable data points is predetermined.

Figure 5.2 shows the information gain from using 3000 data points, using a range of number of clusters from 2 to 40, and two different methods of center-cluster-based choice and random choice. It can be seen clearly in the figure that, given the appropriate choice of clusters, the center-based method works much better. Three clusters is the optimal selection for this data set. However, very high numbers of clusters also show good performances. A random selection without clustering resulted in an information gain of $4.148e + 12$ which is about significantly lower than the optimal information gain through the proposed clustering method.

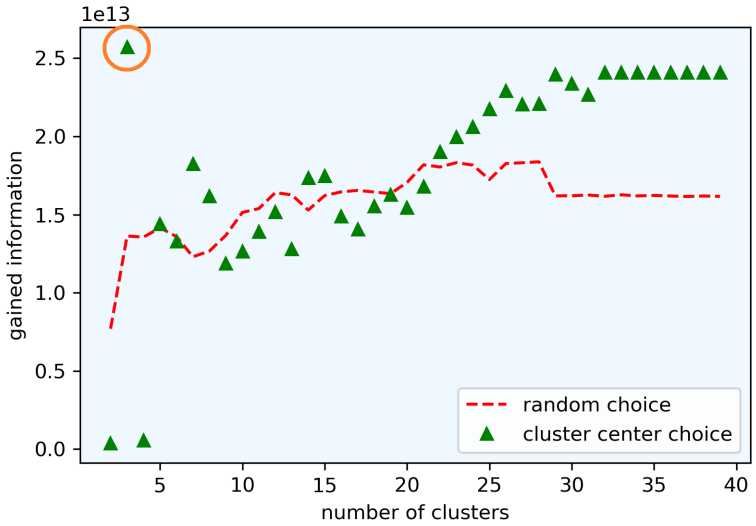

Figure 5.2: Gained information based on the number of clusters and two methods of random selection and cluster-based selection.

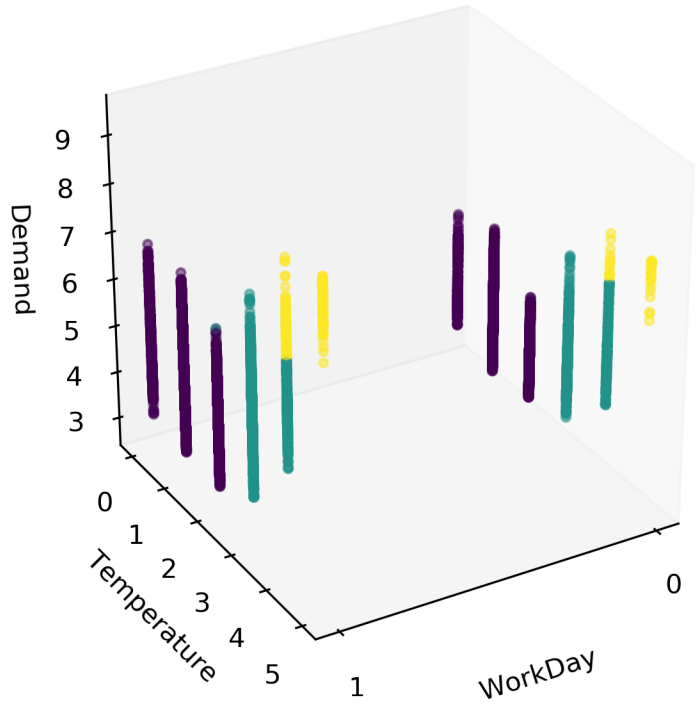

Figure 5.3: The complete *elecdemand* data in three clusters using Birch algorithm. The colors represent the clusters.

5.5 Conclusion

In this work, the problem of data selection from an extremely large data set was investigated. The exact solution of this problem is the optimum point of an MINLP optimization problem which is impossible to solve. Therefore, to preserve all the important and distinguishing properties of the data set, a method based on clustering was proposed. BIRCH clustering algorithm is suggested as the best clustering alternative because of the efficiency of this algorithm in dealing with large data sets. A small numerical study is provided to evaluate the power of the proposed method and compare the available options. The results of the numerical example shows that using Birch clustering

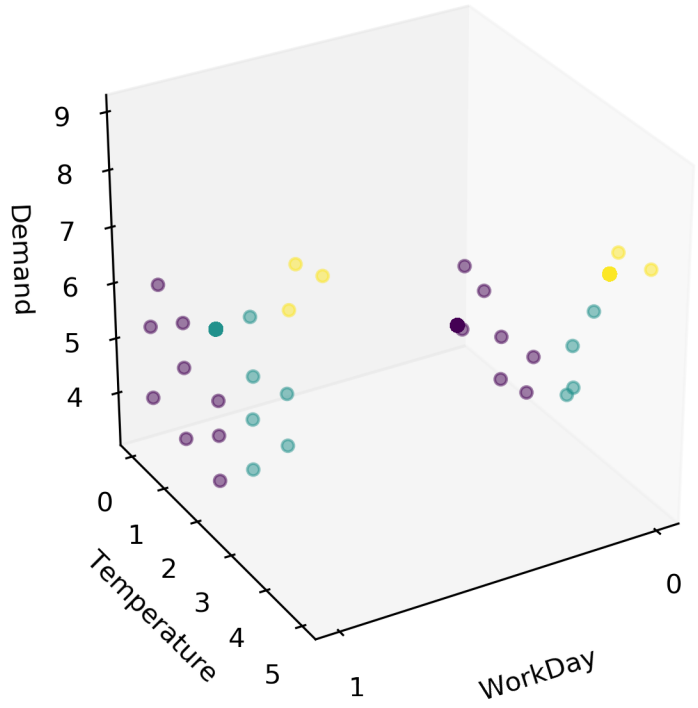

Figure 5.4: The 3000 selected data points from *elecdemand* data using the center-based method. The colors represent the clusters.

algorithm, with center-based data selection and an appropriate number of clusters results in an information gain of much higher than the random selection method with or without clustering. It is noteworthy that random selection from clusters is still much better than a simple random selection. Future research may be focused on a variety of distributions and other statistical regression and classification models. It would be fruitful to find out if there is a global range of clusters that work well, based on the parameters of the number of covariate levels, statistical model, size of original and desired data, etc.

References

Guan, Y., Yang, S., & Zhang, D. H. (2018). Construction of reactive potential energy surfaces with Gaussian process regression: active data selection. *Molecular Physics*, *116*(7-8), 823–834.

Herbold, S. (2013). Training data selection for cross-project defect prediction. *Proceedings of the 9th international conference on predictive models in software engineering*, 1–10.

Hyndman, R. (2020). *fpp2: Data for "Forecasting: Principles and Practice" (2nd Edition)* [R package version 2.4]. https://CRAN.R-project.org/package=fpp2

Kawulok, M., & Nalepa, J. (2012). Support vector machines training data selection using a genetic algorithm. *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*, 557–565.

Khan, N. M., Abraham, N., & Hon, M. (2019). Transfer learning with intelligent training data selection for prediction of Alzheimer's disease. *IEEE Access*, *7*, 72726–72735.

Paudel, S., Elmitri, M., Couturier, S., Nguyen, P. H., Kamphuis, R., Lacarrière, B., & Le Corre, O. (2017). A relevant data selection method for energy consumption prediction of low energy building based on support vector machine. *Energy and Buildings*, *138*, 240–256.

Shanks, D. R. (2017). Regressive research: The pitfalls of post hoc data selection in the study of unconscious mental processes. *Psychonomic Bulletin & Review*, *24*(3), 752–775.

van der Wees, M., Bisazza, A., & Monz, C. (2017). Dynamic data selection for neural machine translation. *arXiv preprint arXiv:1708.00712*.

Zhang, T., Ramakrishnan, R., & Livny, M. (1997). BIRCH: A new data clustering algorithm and its applications. *Data Mining and Knowledge Discovery*, *1*(2), 141–182.

6 Summary

In reliability engineering, numerous methods have been created and utilized based on the special nature of each problem. However, the potential of Phase-type distribution in facing complicated reliability data, its robustness and ability in removing the model selection procedure has prompted this dissertation. Accelerated life testing data, which is a significant product reliability testing method and data source, is investigated in this dissertation and a Phase-type method is proposed and proved the most strong based on the numerical studies. Censored and aggregated data, are among abundantly available in real world reliability data sets. Due to the missing values in these types of data, distribution fitting may not be straight-forward or possible based on the selected probability distribution. ML and Bayesian methods have been proposed based on Phase-type distribution to remove the model selection process while applying a robust data analysis model, using Phase-type distribution. Another research gap is that along with the advances in technology, reliability data sets have grown immensely. In the presence of limited computational resources data selection emerges as an important solution for data analysis. Chapter 5 proposes a data selection method to maximize the information gain from the data.

As the data gathering methods grow, a large portion of data is collected from customers and computers. These types of data, while being invaluable resources, are entitled to error and pointless extra information. Future research can be focused on analysis of error-prone data and covariate selection in the presence of many necessary as well as unnecessary covariates through Phase-type distribution.